UNIVERSIDADE FEDERAL DO ABC

PÓS-GRADUAÇÃO EM NANOCIÊNCIA E MATERIAIS AVANÇADOS

HENRIQUE FERREIRA DOS SANTOS

CIÊNCIA DE DADOS DE MATERIAIS APLICADA À DESCOBERTA DE SEMICONDUTORES

SANTO ANDRÉ - SP

2025

Henrique Ferreira dos Santos

Ciência de Dados de Materiais aplicada à descoberta de Semicondutores

Tese apresentada ao Programa de Pós-Graduação em Nanociência e Materiais Avançados da Universidade Federal do ABC, como requisito parcial para a obtenção do título de Doutor em Nanociência e Materiais Avançados.

Orientador: Prof. Dr. Gustavo Martini Dalpian

Coorientador: Prof. Dr. José Antonio Souza

Santo André

2025

Sistema de Bibliotecas da Universidade Federal do ABC
Elaborada pelo Sistema de Geração de Ficha Catalográfica da UFABC
com os dados fornecidos pelo(a) autor(a).

Ferreira dos Santos, Henrique
Ciência de Dados de Materiais aplicada à descoberta de Semicondutores / Henrique Ferreira dos Santos. — 2025.

290 fls. : il.

Orientação de: Gustavo Martini Dalpian
Coorientação de: José Antonio Souza

Tese (Doutorado) — Universidade Federal do ABC, Programa de Pós-Graduação em Nanociências e Materiais Avançados, Santo André, 2025.

1. Informática de Materiais. 2. Simulação Computacional. 3. Teoria do Funcional da Densidade. 4. Semicondutores. 5. Perovskitas. I. Martini Dalpian, Gustavo. II. Souza, José Antonio. III. Programa de Pós-Graduação em Nanociências e Materiais Avançados, 2025. IV. Título.

Este exemplar foi revisado e alterado em relação à versão original, de acordo com as observações levantadas pela banca examinadora no dia da defesa, sob responsabilidade única do autor e com a anuência do orientador.

MINISTÉRIO DA EDUCAÇÃO
Fundação Universidade Federal do ABC
Avenida dos Estados, 5001 - Bairro Santa Terezinha - Santo André - SP
CEP 09210-580 · Fone: (11) 4996-0017

Ata de Defesa de Tese de Doutorado e Folha de Assinaturas

No dia 18 de Setembro de 2025 às 14:00, no local: online, realizou-se a Defesa da Tese de Doutorado, que constou da apresentação do trabalho intitulado "CIÊNCIA DE DADOS DE MATERIAIS APLICADA À DESCOBERTA DE SEMICONDUTORES" de autoria do candidato, HENRIQUE FERREIRA DOS SANTOS, RA nº 23202010917, discente do Programa de Pós-Graduação em NANOCIÊNCIAS E MATERIAIS AVANÇADOS da UFABC, sob orientação do Profº GUSTAVO MARTINI DALPIAN. Concluídos os trabalhos de apresentação e arguição, o candidato foi considerado APROVADO pela Banca Examinadora.

E, para constar, foi lavrada a presente ata e folha de assinaturas assinada pelos membros da Banca.

gov.br Documento assinado digitalmente
GUSTAVO MARTINI DALPIAN
Data: 18/09/2025 18:31:28-0300
Verifique em https://validar.iti.gov.br

Dr. GUSTAVO MARTINI DALPIAN, USP
Presidente - Interno ao Programa

gov.br Documento assinado digitalmente
PEDRO ALVES DA SILVA AUTRETO
Data: 18/09/2025 20:37:16-0300
Verifique em https://validar.iti.gov.br

Dr. PEDRO ALVES DA SILVA AUTRETO, UFABC
Membro Titular - Examinador(a) Interno ao Programa

gov.br Documento assinado digitalmente
ANDRE SARTO POLO
Data: 18/09/2025 19:43:27-0300
Verifique em https://validar.iti.gov.br

Dr. ANDRE SARTO POLO, UFABC
Membro Titular - Examinador(a) Interno ao Programa

gov.br Documento assinado digitalmente
ROBERTO HIROKI MIWA
Data: 19/09/2025 08:35:22-0300
Verifique em https://validar.iti.gov.br

Dr. ROBERTO HIROKI MIWA, UFU
Membro Titular - Examinador(a) Externo à Instituição

gov.br Documento assinado digitalmente
FELIPE DAVID CRASTO DE LIMA
Data: 18/09/2025 18:40:48-0300
Verifique em https://validar.iti.gov.br

Dr. FELIPE DAVID CRASTO DE LIMA
Membro Titular - Examinador(a) Externo à Instituição

MINISTÉRIO DA EDUCAÇÃO
Fundação Universidade Federal do ABC
Avenida dos Estados, 5001 - Bairro Santa Terezinha - Santo André - SP
CEP 09210-580 · Fone: (11) 4996-0017

Dr. THIAGO BRANQUINHO DE QUEIROZ, UFABC
Membro Suplente - Examinador(a) Interno ao Programa

Dr. RODRIGO BARBOSA CAPAZ, UFRJ
Membro Suplente - Examinador(a) Externo à Instituição

Dr. LUIZ FELIPE CAVALCANTI PEREIRA, UFPE
Membro Suplente - Examinador(a) Externo à Instituição

Dr. ALEXANDRE REILY ROCHA, UNESP
Membro Suplente - Examinador(a) Externo à Instituição

Dra. LUANA SUCUPIRA PEDROZA
Membro Suplente - Examinador(a) Externo à Instituição

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001

Aos meus avós, Armando e Dilza, Anezio e Anna.

Agradecimentos

Todo trabalho, mesmo que executado individualmente, sustenta-se através de elos sociais: nenhum ser humano consegue isolar-se por completo da humanidade ao executar um empreendimento. É, portanto, que não posso deixar de agradecer as diversas pessoas que cruzaram meu caminho ao longo da execução deste trabalho e que exerceram papel fundamental, direta ou indiretamente, para que este fosse concluído. Gostaria de começar agradecendo minha orientadora de mestrado, Ana Melva Champi, que me introduziu na carreira científica formal e foi quem me convidou ao programa de pós-graduação da Nano na UFABC juntamente com a professora Iseli Lourenço Nantes; Ao meu coorientador José Antonio Souza, quem me mostrou o tema de LEDs UV que abordamos nesse trabalho e forneceu apoio essencial às etapas experimentais aqui realizadas. Ao meu orientador, Gustavo Martini Dalpian, quem me introduziu as simulações computacionais DFT e à Informática de Materiais, que completa o tema dessa dissertação, e que, além de todo o apoio técnico, fomentou meu desenvolvimento profissional, foi paciente com os contratempos e possibilitou que todo este esforço tomasse forma definitiva.

Quero agradecer meus colegas de laboratório, professores e colaboradores, com quem pude aprender em diferentes ocasiões, presenciais e remotas: João Nuno Barbosa Rodrigues, James Moraes de Almeida, Amauri Jardim de Paula, Edward Ferreira Almeida Junior, William Espinosa-García, Fernando Pereira Sabino, Carlos Mera Acosta, Naidel Caturello, Gabriel de Miranda Nascimento, Maykon Lemes, Leonardo Quinteira, Eliane Aparecida Morais, Lucas Martin Farigliano, Thiago Puccinelli, Jesus Camilo Diaz Olivella, Pedro Roberto Lopes Vieira, Daniel de Martini Rivera Ferreira, Gabriel Xavier Pereira, Daniel Carlos Santos, Lucas Bandeira e Elton Ogoshi de Melo.

Sou grato as intermináveis conversas com meus colegas de longa data, Tito Spadini, Marcelo Grave, Gabriel Tanuri e Matheus Pessôa, que mesmo não trabalhando diretamente com simulação de materiais, sempre estiverem dispostos a discutir os temas da minha pesquisa e auxiliar com a visão externa de engenheiros de informação, geógrafo e biofísico, respectivamente.

Agradeço aos membros da banca de qualificação Fabio Furlan Ferreira, Roberto Gomes de Aguiar Veiga, Ronaldo Giro e Antonio Claudio Padilha pelas asserções realizadas ao

desenvolvimento deste projeto e que contribuíram de maneira significativa para a versão atual deste texto.

Aos membros da banca de defesa, Pedro Alves da Silva Autreto, André Sarto Polo, Roberto Hiroki Miwa e Felipe David Crasto de Lima, pela discussão dos temas aqui apresentados e sugestões de ajustes finais para prover fluidez e precisão a este trabalho.

Não poderia deixar de agradecer a CAPES pela bolsa de doutorado dentro do projeto de Prevenção e Combate a Surtos, Endemias, Epidemias e Pandemias; Ao Laboratório Nacional de Computação Científica (LNCC) pelo uso do supercomputador Santos Dumont; A Central Computacional Multiusuário (CCM) e a Central Experimental Multiusuário (CEM) da UFABC pelo uso, respectivamente, dos recursos computacionais e equipamento de caraterização que foram empregados ao longo deste trabalho.

Também aproveito para agradecer a Sandra Regina Rodrigues Ribeiro, Rosana Batista Gimenes Biz, Luiz Carlos dos Santos e Marcelo Meireles dos Santos, do Departamento de Física dos Materiais e Mecânica do Instituto de Física da Universidade de São Paulo (IFUSP) onde realizei parte desse doutorado, na sala 205 do Alessandro Volta.

Completo agradecendo aos meus pais, Armando e Silvana, e ao meu irmão, Victor Gabriel, cujo suporte familiar foi importantíssimo para que esta tarefa fosse concluída.

Obrigado a todos!

É possível alcançar por certo, em relativamente pouco tempo, determinado êxito, apropriando-se de técnicas e truques; mas tais resultados permanecem superficiais e insatisfatórios, pois deixam sempre o estudante desamparado em situações novas e inesperadas. Se ele não foi treinado a ter uma perspectiva do desenvolvimento orgânico, nenhum conhecimento e aplicação de motivos modernos, por mais elaborados que sejam, podem capacitá-lo a realizar um trabalho criativo. - Walter Gropius, 1937.

Resumo

Ciência de Dados de Materiais é definida pela intersecção entre a Informática de Materiais e a Simulação Computacional de Materiais, visando a aplicação de técnicas de Inteligência Artificial e Aprendizado de Máquina para a Ciência de Materiais. Um dos seus principais objetivos é fornecer um paradigma para o design inverso de materiais, aprimorando a descoberta de novos compostos. Nesta tese utilizamos este tipo de metodologia na busca por materiais que tenham um gap de energia grande, objetivando a descoberta de potenciais candidatos para LEDs que emitam na região do ultravioleta. O potencial desta metodologia é demonstrado através de uma abordagem hierárquica de metodologias em diferentes cenários: no primeiro foi usada a abordagem convencional de experimentos e simulações utilizando Teoria do Funciona da Densidade (DFT) para estudar perovskitas híbridas pseudo-2D do tipo Dion-Jacobson; no segundo, foram realizados cálculos *high throughput* DFT para estudar uma nova família de compostos perovskitas pseudo-1D obtidos através de substituição química na estrutura protótipo da jakobssonita, um mineral vulcânico de composição $CaAlF_5$; no terceiro, gerou-se compostos binários usando combinações químicas com a subsequente exploração desse espaço por modelos de aprendizado de máquina treinados com *features* atômicas para selecionar semicondutores estáveis. Esses estudos correspondem a uma gradativa mudança de paradigma, com cerca de 5 variações de materiais no primeiro caso, cerca de 3 mil materiais no segundo, e mais de 1 milhão no terceiro. Os resultados desta tese evidenciam a robustez da metodologia proposta, estabelecendo novas fronteiras e possibilidades para o desenvolvimento de materiais com propriedades direcionadas e aplicadas a uma funcionalidade específica.

Palavras chave: ciência de dados de materiais, informática de materiais, simulação computacional, teoria do funcional da densidade, semicondutores, jakobssonita, perovskita.

Abstract

Materials Data Science is defined by the intersection between Materials Informatics and Computational Materials Simulation, aiming at the application of Artificial Intelligence and Machine Learning techniques to Materials Science. One of its main objectives is to provide a paradigm for the inverse design of materials, thereby enhancing the discovery of new compounds. In this thesis, we employ this methodology in the search for materials with large band gaps, targeting the discovery of potential candidates for ultraviolet-emitting LEDs. The potential of this methodology is demonstrated through a hierarchical approach across different scenarios: in the first, the conventional strategy of experiments and Densistyt Functional Theory (DFT) simulations was used to study pseudo-2D Dion–Jacobson hybrid perovskites; in the second, high-throughput DFT calculations were performed to investigate a new family of pseudo-1D perovskite compounds obtained via chemical substitution in the prototype structure of jakobssonite, a volcanic mineral with composition $CaAlF_5$; in the third, binary compounds were generated from chemical combinations, and this space was subsequently explored using machine learning models trained on atomic features to select stable semiconductors. These studies reflect a gradual paradigm shift: around 5 material variations in the first case, about 3 thousand in the second, and more than 1 million in the third. The results of this thesis highlight the robustness of the proposed methodology, establishing new frontiers and possibilities for the development of materials with tailored properties, applied to a specific functionality.

Keywords: materials data science, materials informatics, computational materials, density functional theory, semiconductors, jakobssonite, perovskite.

Conteúdo

1 Introdução **19**

1.1 Motivação 19

1.2 Objetivos 27

1.3 Estrutura do Trabalho 27

2 Abordagem Convencional **31**

2.1 Simulação Computacional de Materiais 32

2.1.1 Estrutura Eletrônica 33

2.1.1.1 Método de Hartree-Fock 35

2.1.1.2 Teoria do Funcional da Densidade 37

2.1.1.3 Métodos Híbridos 43

2.1.1.4 Teoria de Perturbação de Muitos Corpos 44

2.1.1.5 Teoria do Funcional da Densidade Dependente do Tempo . 45

2.1.1.6 Softwares para Métodos Quânticos 46

2.1.2 Dinâmica Molecular 47

2.2 Estabilidade Termodinâmica de Materiais 50

2.2.1 Entalpia de Formação 50

2.2.2 Convex Hull 51

2.2.3 Estabilidade Dinâmica 52

2.2.4 Outras Estabilidades 52

2.3 Síntese e Caraterização de Materiais 53

2.3.1 Rotas de Síntese 54

2.3.2 Técnicas de Caracterização 56

3 Ciência de Dados de Materiais **61**

3.1 Fundamentos de Inteligência Artificial 63

3.1.1 Inteligência Artificial Simbólica 65

3.1.1.1 Problemas de Busca 66

3.1.1.2 Explicabilidade e Interpretabilidade 69

3.1.2 Aprendizado de Máquina 69

3.1.2.1 Algoritmos de Aprendizado de Máquina 71

3.1.2.2 Representação de Dados 74

3.1.2.3 Dados Sintéticos e Aumento de Dados 75

3.1.2.4 Avaliação de Desempenho 76

3.1.2.5 Meta-Aprendizado 80

3.2 Informática de Materiais 81

3.2.1 Simulações High Throughput 81
3.2.2 Bases de Dados 84
3.2.3 Ontologias e Metadados 85
3.2.4 Mineração de Dados 88
3.2.4.1 Screening de Materiais 89
3.2.4.2 Mineração de Texto 89
3.3 Inteligência Artificial em Ciência dos Materiais 92
3.3.1 Representação de Dados 92
3.3.2 Interpretabilidade e Explicabilidade 94
3.3.3 Modelos Substitutos 96
3.3.4 Modelos Generativos 96
3.3.5 Laboratórios Autônomos 98

4 Estudo de Caso I: Perovskitas Híbridas pseudo-2D Dion-Jacobson 101
4.1 Metodologia 105
4.1.1 Metodologia Experimental 105
4.1.2 Metodologia Computacional 106
4.2 Resultados e Discussão 107
4.3 Conclusão 119

5 Estudo de Caso II: Jakobssonita e Perovskitas pseudo-1D 121
5.1 Metodologia 122
5.1.1 Espaço de Busca 122
5.1.2 Simulações Computacionais HT-DFT 125
5.2 Resultados e Discussão 126
5.2.1 Screening e Aumento de Dados 126
5.2.2 Estabilidade Termodinâmica 130
5.2.3 Propriedades Eletrônicas 133
5.2.4 Seleção 136
5.2.5 Propriedades Óticas 138
5.3 Conclusão 139

6 Estudo de Caso III: Semicondutores Binários 141
6.1 Metodologia 143
6.1.1 Datasets 143
6.1.2 Aprendizado de Máquina 146
6.1.2.1 Algoritmos 146
6.1.2.2 Features 146
6.1.2.3 Avaliação de Desempenho 147
6.2 Resultados e Discussão 148
6.2.1 Análise Exploratória 148
6.2.2 Construção dos Modelos de Predição 161

6.2.3 Interpretabilidade dos Modelos 163
6.2.4 Geração de Novos Materiais 167
6.3 Conclusão 169

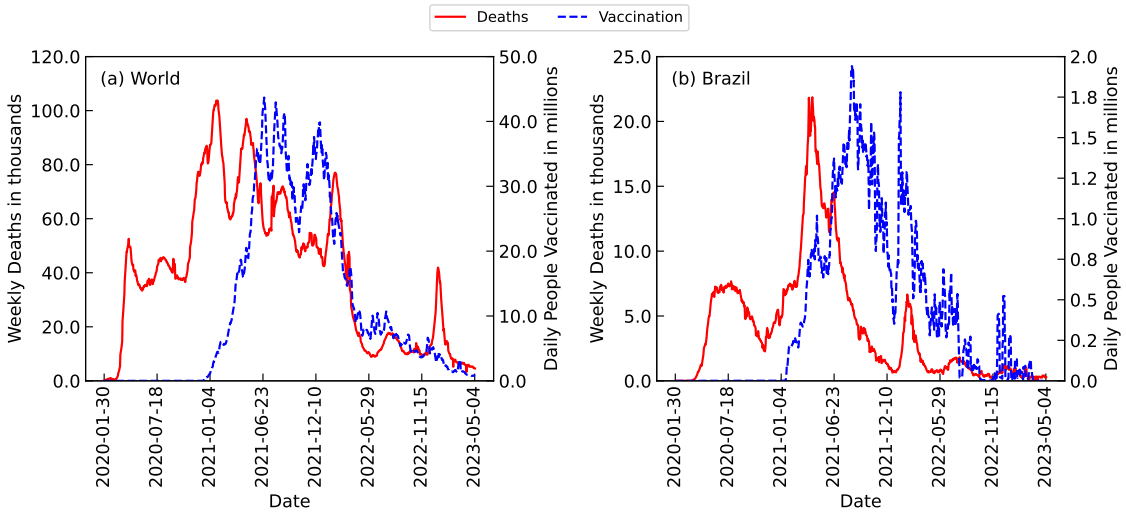
7 Conclusões **171**

1 Introdução

1.1 Motivação

Com a pandemia de COVID19 causada pelo vírus SARS-CoV-2 (coronavírus 2 da Síndrome Respiratória Aguda Grave, do inglês, Severe Acute Respiratory Syndrome Coronavirus 2) [1], a humanidade enfrentou uma crise de saúde pública sem precedentes na era globalizada. No dia 30 de janeiro de 2020, a Organização Mundial da Saúde (OMS) declarou Emergência de Saúde Pública de Importância Internacional, cujo fim foi anunciado somente em 5 de maio de 2023 [2]. No mundo, nestes 39 meses, morreram 6.928.581 pessoas devido a covid, sendo no Brasil cerca de 701.494 [3]. O avanço científico e tecnológico, aliado às medidas de saúde pública adotadas por diferentes governos, garantiram o controle e posterior combate dessa doença. No primeiro ano, foram tentadas mais de noventa vacinas diferentes, com pesquisas de vacinas virais, vacinas de vetores virais, vacinas baseadas em proteínas e vacinas de ácido nucleico [4]. Como ilustração, na Figura 1.1 podemos ver os efeitos da vacinação em massa com as curvas de mortalidade declinando assim que as curvas de vacinação subiram quando as primeiras vacinas aprovadas em testes clínicos foram distribuídas.

Figura 1.1: Série temporal do número de mortes semanais devido a SARS-CoV-2 (vermelho) e do número de pessoas vacinadas diariamente (azul) para (a) mundo e (b) Brasil.



Fonte: elaborado pelo autor com dados de [3].

Ao mesmo tempo que medidas emergenciais tiveram que ser tomadas para o combate à pandemia, novos conceitos e tecnologias foram e estão sendo promovidos para aperfeiçoar as técnicas e ferramentas disponíveis a enfrentamentos de pandemias futuras. O desen-

volvimento dos materiais que são utilizados nos produtos e equipamentos de saúde clinica ou diagnóstica é um exemplo. A nanotecnologia desempenha um papel fundamental na criação de biossensores avançados, permitindo a detecção rápida e precisa de agentes patogênicos como o vírus da raiva (*rabies virus*) [5]. Por outro lado, dispositivos de desinfecção utilizam irradiação ultravioleta para eliminar microrganismos a fim de controlar possíveis infecções nosocomiais [6]. Tais dispositivos podem ser construídos com diodos emissores de luz (LEDs, do inglês, light-emitting diodes) cujos materiais devem ser devidamente projetados para emitir nessa faixa do espectro eletromagnético [7].

Diante desse cenário de avanços tecnológicos no combate às pandemias, especialmente no tocante ao desenvolvimento de novos materiais, queremos destacar aquelas relacionadas a produção de luz ultravioleta.

Por definição, o espectro ultravioleta é separado em faixas de comprimento de onda, conforme apresentado na Tabela 1.1. A irradiação ultravioleta germicida (UVGI, do inglês ultraviolet germicidal irradiation) é aquela com comprimento de onda entre 200 e 320 nm [6]. A região do espectro com atividade actínica refere-se aos efeitos fotoquímicos e fotobiológicos da luz, enquanto a região eritemal refere-se à possibilidade de queimaduras na pele devido a exposição à radiação. A região do ultravioleta de vácuo VUV (do inglês, vaccum ultraviolet) é fortemente absorvida pelo oxigênio molecular, O_2, do ar, sendo transparente somente para o nitrogênio molecular, N_2, entre 150-200 nm. Essa faixa também engloba porções ionizantes da radiação, especialmente no denominado ultravioleta extremo EUV (do inglês, extreme ultraviolet) de 10-121 nm.

Tabela 1.1: Faixas do Espectro Ultravioleta.

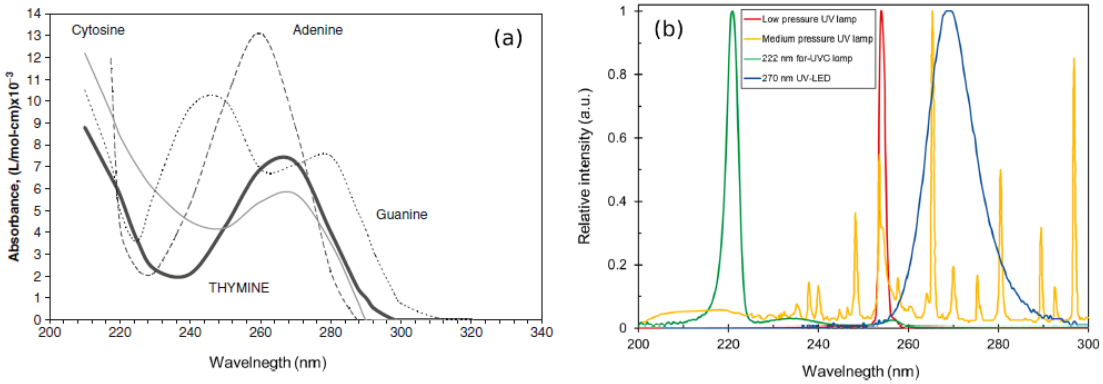
<table>
<tr><th>Faixa</th><th>Comprimento de onda λ (nm)</th><th>Energia do fóton E (eV)</th><th colspan="3">Atividade</th></tr>
<tr><td>UVA</td><td>320 - 400</td><td>3.9 - 3.1</td><td colspan="3">Luz negra</td></tr>
<tr><td>UVB</td><td>280 - 320</td><td>4.4 - 3.9</td><td>Eritemal</td><td rowspan="2">Germicida</td><td rowspan="3">Actínica</td></tr>
<tr><td>UVC</td><td>200 - 280</td><td>6.2 - 4.4</td><td>Produção de O_3</td></tr>
<tr><td>VUV</td><td>10 - 200</td><td>124 - 6.2</td><td colspan="2">Absorvido pela atmosfera</td></tr>
</table>

Fonte: adaptado de [6].

A ação germicida da luz ultravioleta pode ser explicada através do dano causado pela irradiação às cadeias de RNA (ácido ribonucléico) e DNA (ácido desoxirribonucléico) [6, 8, 7] contidas nas células (bactérias) ou dentro do capsídeo viral (vírus). Este dano causado ao material genético do microrganismo ou do vírus interfere na sua capacidade de reprodução, o que denomina-se de inativação [8]. Uma das formas desse dano ocorrer é pela quebra das ligações entre as bases nitrogenadas do ácido nucleico, que passam a formar dímeros na mesma fita [6, 8]. Fisicamente, observa-se que esse processo é favorecido na região do ultravioleta devido a absorção dessa faixa do espectro por essas bases nitrogenadas

(adenina, citosina, guanina e timina), como apresentado na Figura 1.2 (a). Além disso, a dose de irradiação, D, é um fator importante para a completa inativação, sendo ela determinada pelo produto da irradiação, I_r, e o tempo de exposição, t, $D(t) = I_r t$ [7]. Para cada microrganismo ou vírus, em cada situação de desinfecção, há uma dose de irradiação recomendada para inativação mais efetiva. Por exemplo, para o SARS-CoV-2 em superfície sólida, uma dose de $D = 1\,\mathrm{J\,cm^{-2}}$ com comprimento de onda $\lambda = 254\,\mathrm{nm}$ foi capaz de eliminar 90% do vírus [9], enquanto que uma dose de $D = 0.04\,\mathrm{J\,cm^{-2}}$ do mesmo comprimento de onda foi suficiente para alcançar o mesmo grau de inativação quando irradiada sobre mesmo vírus em meio aquoso [8].

Figura 1.2: (a) Espectro de absorção de bases nitrogenadas; (b) Espectro de emissão de diferentes dispositivos UVGI.

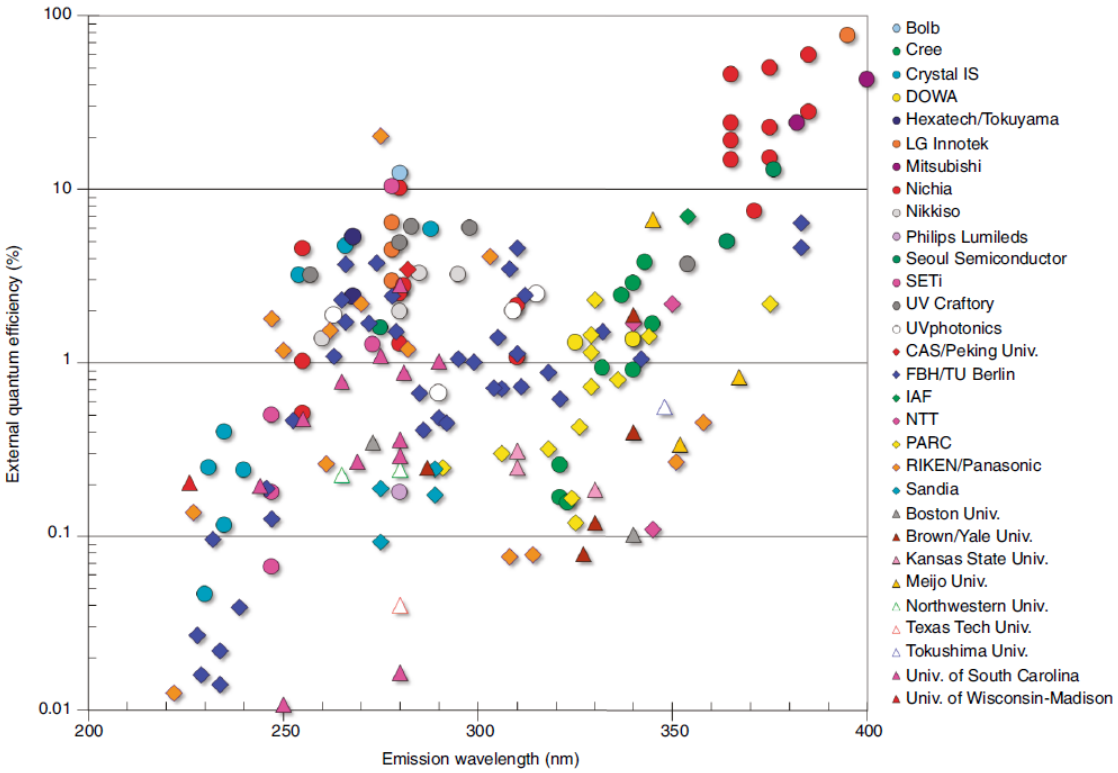


Fonte: (a) extraído de [6]; (b) extraído de [8].

Diferentes equipamentos geram diferentes qualidades de emissão eletromagnética no espectro do ultravioleta. Na Figura 1.2 (b) são apresentados espectros de emissão de quatro diferentes dispositivos. Lâmpadas pulsadas de arco de xenônio e lâmpadas de mercúrio são lâmpadas não polarizadas de baixa e média pressão com dois eletrodos, normalmente mais baratas, mas que possuem materiais tóxicos (mercúrio) ou exigem circuitos eletrônicos adicionais complexos (pulsadas). LEDs, por sua vez, geram emissões mais próximas do máximo de efetividade germicida (de acordo com o espectro de absorção), são mais compactos, normalmente não usam materiais tóxicos e requerem circuitos mais simples. Contudo, seus custos de produção são maiores uma vez que a tecnologia atual é baseada em processos de fabricação de semicondutores [10, 11], como nanoimpressão litográfica [12]. Para critério de comparação, um LED UV CUD7QF1A da Seoul Viosys, com emissão em 275 nm e 1.6 mW de potência, custa U$ 2.79 no mercado internacional, chegando a R$70.00 no mercado brasileiro; uma lâmpada de vapor de mercúrio de Philips TUV TL Mini FAM/10X25BOX com emissão em 253.7 nm e 6 W de potência, custa cerca de U$20.00 no mercado internacional e R$100.00 no Brasil (cotações de outubro de 2023). Portanto, considerando a instalação de potência, o custo total de um aparelho de desinfecção produzido com LEDs tenderá a ser maior do que aquele com lâmpadas de mercúrio.

O desenvolvimento de LEDs emissores no ultravioleta também está sendo limitado pela baixa eficiência quântica externa $\eta_{EQE} = (N_e/t)/(N_\nu/t)$ (do inglês, *External Quantum Efficiency*) que esses dispositivos apresentam, conforme avançam na faixa do UVB e UVC. É importante notar que η_{EQE} está diretamente relacionado à eficiência energética do dispositivo, ou seja, a quantidade de energia elétrica necessária (medida em elétrons por segundo, N_e/t) para produzir certa quantidade de fótons por segundo (N_ν/t). Kneissl *et al.* [7] produziram um revisão sobre o estado da arte desses dispositivos, tanto comerciais quanto experimentais: na Figura 1.3 observamos que, para emissões com $\lambda < 365\,\text{nm}$, a η_{EQE} fica menor do que 10%, ou até mesmo próximo de apenas 1% para a região de maior efetividade germicida. Essa transição abrupta na eficiência pode ser atribuída aos materiais usados na produção de cada tipo de LED: é exatamente nesta faixa que ocorre a transição de LEDs produzidos com InGaN para aqueles produzidos com AlGaN [7]. A alteração do material ocorre devido ao gap de energia, uma banda proibida de energia entre os estados eletrônicos de valência e de condução, propriedade esta que determina majoritariamente o comprimento de onda dos fótons que podem ser emitidos pelo dispositivo.

Figura 1.3: Eficiência Quântica Externa de diferentes LEDs ultravioleta.

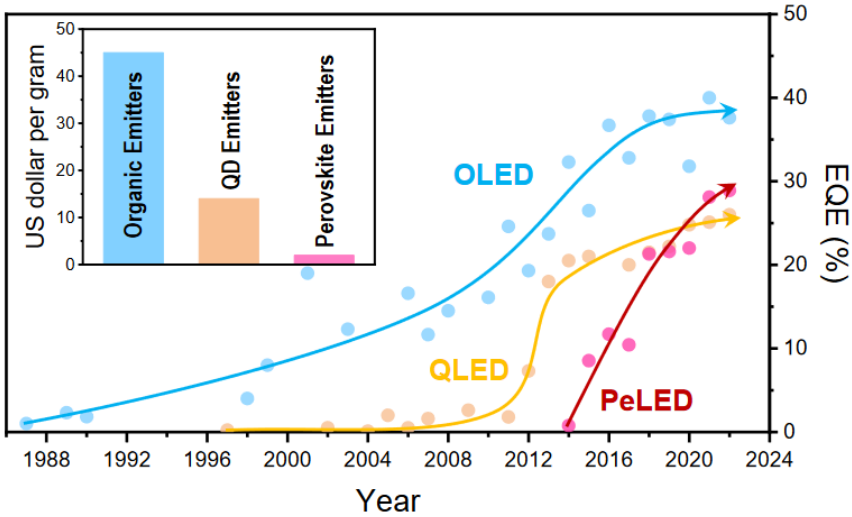


Fonte: extraído de [7].

Os diodos emissores de luz atuais são, em sua maioria, produzidos com materiais semicondutores, em tecnologia de iluminação de estado sólido (do inglês, *solid-state lighting*, SSL) [13]. O fenômeno físico envolvido no processo de geração de fótons neste dispositivos

é denominado eletroluminescência, e consiste na recombinação radiativa de elétrons na banda de condução com buracos eletrônicos na banda de valência [14]. O taxa de fótons gerada depende da quantidade de portadores de carga presentes no dispositivo, de forma que, para controlar essas quantidades, os semicondutores são dopados com impurezas que aumentam a concentração de buracos (semicondutor tipo P) ou de elétrons (semicondutor tipo N). Quando o dispositivo é feito com um único material semicondutor dopado, ele é denominado LED de homojunção PN; quando feito com dois semicondutores diferentes, de heterojunção PN [13]. Também podem ser incorporados efeitos de confinamento quântico na estrutura do material, por exemplo, ao construir a junção com filmes semicondutores de espessura nanométricas (dispositivos denominados de poços quânticos, do inglês *quantum wells*, QW) - o confinamento quântico modifica o valor de bandgap, podendo também influenciar na eficiência. O estado da arte em LEDs emissores no UVC é exatamente a tecnologia de semicondutores (AlGaN) dopados, com heterojunção, construídos com múltiplas camadas de poços quânticos (*multiple quantum well structure*, MQW) [7]. Como alternativa tecnológica, existem, por exemplo, os LEDs baseados em perovskitas [15], PeLEDs. Estes dispositivos possuem rotas de fabricação mais baratas e também podem ser estruturados para apresentar confinamento quântico (como QW [16], ou como pontos quânticos, do inglês, *quatum dots*, QD [17]), entretanto não são suficientemente estáveis e se concentram, em sua maioria, em emissões dentro do espectro visível. Apesar disso, as pesquisas em PeLEDs tem avançado consistentemente nos últimos anos, como mostra a Figura 1.4:

Figura 1.4: Eficiência Quântica Externa de diferentes tecnologias de LEDs: orgânicos (OLED), pontos quânticos (QLED) e de perovskita (PeLED); *inset* com o custo do principal percursor utilizado na síntese.



Fonte: extraído de [18].

A rota de aumento de performance desses dispositivos, diminuindo seu custo de produ-

ção e aumentando sua eficiência energética e estabilidade, passa pela descoberta de novos materiais e processos de fabricação. Um paralelo interessante é o desenvolvimento comercial da lâmpada elétrica incandescente no século XIX, no que ficou conhecida como *abordagem edsionina*, devido a Thomas Edison: testar diferentes materiais e métodos sistematicamente por "tentativa e erro". Apesar da eficácia dessa abordagem, sua eficiência é questionável, principalmente se considerados os custos, tanto de tempo quanto de dinheiro, envolvidos na experimentação. Por outro lado, o desenvolvimento teórico e computacional estabeleceu alicerce para um *design racional*, onde cálculos e experimentos *in silico* permitem orientar o processo de descoberta, reduzindo o espaço de busca da solução [19]. Aliar essas abordagens, primeiro orientando racionalmente a descoberta, e depois testando somente os candidatos com maior probabilidade de sucesso, torna-se um caminho natural.

Além disso, o avanço do poder computacional, em especial quanto ao armazenamento de dados e a intercomunicação entre dispositivos, tem incrementado radicalmente as possibilidades para uma racionalização da descoberta científica e tecnológica. Observando o desenvolvimento da Bioinformática no final do século XX e primeira década do século XXI, podemos entender o contexto da Ciência de Materiais atual: o sequencialmente do código genético humano objetivado pelo Projeto Genoma Humano (HGP) exigiu a criação de infraestrutura computacional para processamento e armazenamento de grandes volumes de dados [20]; por outro lado, para a área de materiais, o que analogamente se convencionou chamar de Informática de Materiais, temos o exemplo da Iniciativa Genoma de Materiais (do inglês, *Materials Genome Initiative*), lançada em 2011 pelo governo norte-americano que a descreve como:

> "[...] um esforço conjunto de várias agências federais voltado para a descoberta, produção e implementação de materiais avançados em metade do tempo e a uma fração do custo quando comparado aos métodos tradicionais."[21]

Ambos os projetos são exemplos de um novo paradigma de se fazer ciência, a Ciência Orientada por Dados (do inglês, *Data-Driven Science*) [19]. Neste paradigma, técnicas computacionais são aplicadas não só para modelagem e simulação, porém também para um tratamento intensivo de dados, desde a coleta, o armazenamento, a distribuição e a recuperação da informação contida neles, visando a curadoria e a descoberta de novos padrões, isto é, a extração de novas informações e conhecimentos úteis. Este paradigma contrasta com o da Ciência Orientada por Hipóteses, onde primeiro são feitas perguntas específicas e depois coletados os dados para confirmá-las ou negá-las [22]. Na área de materiais, é neste contexto que surgem bases de dados abertas de resultados de simulações computacionais como AFLOW, OQMD, Materials Project e NOMAD [23]. Além da organização e distribuição, a ciência orientada por dados se vale de técnicas tradicionalmente pertencentes à Ciência da Computação, como a mineração de dados, a mineração de texto e a visão computacional para expandir a forma como a informação é recuperada, não só de acervos devidamente projetados (como as bases de dados anteriormente citadas), como também de acervos de escopo mais amplo como toda a literatura científica, seja na forma

de texto, de imagens ou de gráficos. Mais do que isso, essas novas ferramentas auxiliam na descoberta de padrões que emergem dessa grande quantidade de dados e que não são facilmente observáveis. Em especial, o uso de técnicas de inteligência artificial e aprendizado de máquina tem estabelecido novas fronteiras de pesquisa nas mais variadas disciplinas científicas, como Ciência de Materiais [19, 23, 24, 25], Física [26, 27] e Biologia [28]. Não obstante aos dois projetos anteriormente apresentados, é evidente que a abordagem orientada a dados também foi concomitantemente desenvolvida e aperfeiçoada em outros domínios além da Bioinformática e da Informática de Materiais, como a Astronomia[29] e a Física de Altas Energias[30], onde instrumentos como telescópios, satélites e aceleradores de partículas geram grandes quantidades de dados diariamente. Sem o uso de ferramentas computacionais propriamente projetadas para trabalhar com esse grande volume de dados, não seria possível extrair ciência dessa quantidade de informação. Deste mesmo modo, tais conceitos e ferramentas mostram-se transdisciplinares, podendo ser usados por todas as disciplinas do conhecimento, e mesmo assim, antedisciplinares[1] [31] ao estabelecerem uma nova área do conhecimento, a Ciência de Dados.

Neste contexto, estabelecer um novo material para a construção de LEDs UV deve passar por um uso intensivo dos dados já estabelecidos, por exemplo, através de simulações computacionais de materiais disponíveis nas bases de dados; também deve passar por um forte embasamento teórico, assistido por novas simulações que complementem as lacunas percebidas; finalmente, essas etapas de racionalização precedentes devem diminuir o espaço de soluções a serem testadas empiricamente no laboratório. Em se tratando da síntese de um novo material, essa metodologia deve portanto estabelecer a *sintetizabilidade*[2] do mesmo, isto é, se o material pode de fato existir nas condições esperadas; essa metodologia também deve determinar as propriedades de interesse para a aplicação, por exemplo, se o material é capaz de emitir luz com comprimento de onda $\lambda \approx 265\,\mathrm{nm}$[3], onde há o pico de efetividade germicida. Dessa forma, por um lado será determinada a possibilidade da solução e por outro, se de fato a solução é aplicável. Mais do que isso, o acréscimo de critérios racionais e heurísticos, com a devida ordenação das soluções descobertas, permite uma etapa final de otimização, onde não só soluções reais e aplicáveis são elencadas, como também, ranqueadas. Essa otimização pré-empírica visa diminuir os custos financeiros e de tempo no aperfeiçoamento tecnológico, aumentando não só a eficiência, mas a efetividade desses esforços.

Uma vez que definimos nosso problema (descoberta de novos materiais com exemplo de aplicação na construção de LEDs UV para desinfecção) e apresentamos as diretrizes da nossa metodologia (que aqui denominaremos de Ciência de Dados de Materiais), é oportuno destacar que a motivação introduzida não deve se ater apenas ao *solucionismo*

[1] Neste texto usamos esse termo como expressão do processo que antecede a criação de uma disciplina, ou seja, quando uma área de conhecimento nova ainda não foi incorporada como disciplina autônoma.

[2] Aqui estamos usando o neologismo *sintetizabilidade*, do inglês *synthesizability*, para expressar a ideia de que um material pode ser obtido por uma rota de síntese artificial.

[3] Os principais fatores utilizados para escolher um material no caso das tecnologias de LED atuais são o tamanho do bandgap, condições de casamento entre os parâmetros de rede para o crescimento epitaxial e a facilidade de dopagem do tipo p [14].

tecnológico, isto é, a crença de que problemas complexos (como pandemias) podem ser reduzidos a soluções de engenharia que produzam artefatos tecnológicos (como LEDs). É evidente que tais aparatos são elementos importantes para atacar esses problemas, porém soluções reais envolvem dimensões de esferas maiores, como as econômicas, as sociais e as políticas. Os resultados aqui apresentados e discutidos devem ser entendidos não só como elementos necessários a inovação tecnológica, mas, também, e sobretudo, como subsídios a conhecimentos que podem e devem ser expandidos para se tornarem soluções de fato.

Por fim, é oportuno dizer que a Ciência, em especial as Ciências Naturais e de Engenharia, sempre foram orientada por dados. Quando chamamos a atenção para a mudança de paradigma, na verdade estamos chamando atenção para dois fatores importantes: o primeiro relacionado às novas ferramentas e o segundo relacionado a compartimentação do conhecimento (disciplinaridade).

Quanto ao primeiro, convém citar um pensamento do físico Freeman Dyson:

> "Novas direções na ciência são criadas por novas ferramentas com muito mais frequência do que por novos conceitos. O efeito de uma revolução impulsionada por conceitos é explicar coisas antigas de novas maneiras. O efeito de uma revolução impulsionada por ferramentas é descobrir coisas novas que precisam ser explicadas" – Freeman Dyson, *Imagined Worlds*, 1997. [32]

E quanto ao segundo ponto:

> "O progresso é impulsionado por novas questões científicas, que exigem novas formas de pensar. Você quer ir para onde uma pergunta te leva, não para onde seu treinamento te deixou." – Sean Eddy, *Antedisciplinary Science*, 2004. [31]

Enquanto estes dois pontos são reflexões de caráter epistemológico, no sentido de que abordam como o conhecimento científico é produzido e transformado, podemos levantar um terceiro ponto que traz uma perspectiva histórica da aplicação de métodos teóricos para as concepções tecnológicas. O matemático Nobert Wiener comenta sobre o processo de eletrificação de plantas industriais, anteriormente movidas por máquinas à vapor - que requeriam um alto grau de sofisticação mecânica, que estava distante das leis físicas, no sentido de serem soluções *ad hoc* oriundas da experiência de seus projetistas:

> "A construção de máquinas que envolvam tais partes foi transferida do domínio do artífice especializado para o do pesquisador de laboratório, e nesse campo ele dispõe de todos os utensílios disponíveis da teoria dos circuitos para substituir uma inventiva mecânica do antigo grupo. A invenção, no sentido antigo, foi suplantada pela utilização inteligente de certas leis da Natureza. O passo das leis da Natureza para a sua utilização foi reduzido de uma centena de vezes. " – Nobert Wiener, *The Human Use of Human Beings*, 1954. [33]

Nenhum desses autores tinha em mente a Ciência de Dados de Materiais ao elaborarem esses pensamentos. Contudo, tais ideias destacam o cerne de se estabelecer novas

abordagens dentro da área de materiais para gerar avanços científicos e tecnológicos. As ferramentas e conceitos, oriundos em sua maioria da computação, têm apresentado uma forte contribuição para essa área numa progressão (multidisciplinaridade, interdisciplinaridade, transdisciplinaridade e antedisciplinaridade) que necessita de um novo nome - a Ciência de Dados de Materiais. A automação, tanto na geração de dados quanto no processo criativo, permite encontrar padrões antes não vistos, contribuindo com novas percepções, uma maior racionalização do design de materiais e dispositivos, e, eventualmente, na descoberta de novos fenômenos. Isso contribuiu para a tendência de tornar os modelos teóricos - as leis da natureza - diretamente aplicáveis nos processos de inovação, reduzindo o hiato entre formulação científica e aplicação tecnológica.

1.2 Objetivos

Vista a delimitação da nossa motivação geral, gostaríamos de enfatizar os dois objetivos centrais que foram almejados no desenvolvimento deste trabalho:

- Estudar métodos para a descoberta de novos materiais, discutindo as limitações e oportunidades desse novo paradigma, a Ciência de Dados de Materiais, tanto em suas diferentes metodologias internas, quanto na sua inserção nas linhas de pesquisa e desenvolvimento já estabelecidas;

- Gerar listas de materiais candidatos para uma aplicação de exemplo determinada e que tal tarefa seja feita de maneira racional. Como aplicação de exemplo focamos em semicondutores de gap ultra-largo, como discutido na motivação, porém sem, necessariamente, enfatizarmos as particularidades da construção de LEDs (isto é, figuras de mérito específicas desses dispositivos).

1.3 Estrutura do Trabalho

No presente capítulo (Capítulo 1), detalhamos os elementos de motivação para a escolha dos dois principais objetivos deste trabalho: o estudo de novas abordagens no design e descoberta de materiais com exemplo de aplicação na geração de listas de materiais candidatos para construção de LEDs UV.

No Capítulo 2 tratamos dos fundamentos teóricos utilizados nas abordagens convencionais. Por abordagens convencionais designamos os paradigmas experimental, teórico e computacional tradicionalmente estabelecidos nas Ciências, em contraste com o paradigma dito orientado por grandes volumes de dados (*data-driven*) que é explorado separadamente do Capítulo 3. Nesse estabelecemos um panorama geral das técnicas utilizadas nas novas abordagens de design e descoberta de materiais, que denominamos Ciência de Dados de Materiais.

Os Capítulos 4, 5 e 6 constituem três estudos de caso da aplicação das metodologias discutidas nos Capítulos 2 e 3. Estes exemplos foram ordenados de acordo com o grau de

mudança do paradigma utilizado.

No Capítulo 4 apresentamos majoritariamente elementos das técnicas convencionais aplicadas no estudo de perovskitas híbridas metal-orgânica (HOIP) de haletos, com estrutura pseudo-2D do tipo Dion–Jacobson, composições (HDA)$PbBr_4$, (HDA)(MA)Pb_2Br_7 e (HDA)$(MA)_2Pb_3Br_{10}$. Determinamos experimentalmente e computacionalmente o bandgap desses materiais e a energia de ligação de excítons, discutindo aspectos do confinamento quântico morfológico e dielétrico desses materiais de baixa dimensionalidade. Nossos resultados sugerem que quando o número de camadas de octaedros de $PbBr_6$ aumenta, o bandgap diminui e surgem distorções na rede que causam Efeito Rashba. Do ponto de vista metodológico, essa etapa forneceu um exemplo nítido da iteração tradicional de experimento seguido de modelagem computacional. Este tipo de abordagem favorece um ganho de informação em profundidade, isto é, os métodos teóricos e experimentais se revezam progressivamente para que seja gerado conhecimento consistente sobre o material em questão. Em geral, essa abordagem é limitada em largura, focando em apenas uma ou poucas composições e estruturas.

No Capítulo 5 demonstramos o uso de cálculos de DFT *high througtput* para gerar uma família de novos materiais com base na Jakobssonita, um mineral vulcânico de composição $CaAlF_5$. Foram simulados um total de 2933 materiais, que incluíram as 100 estruturas alvo baseadas na jakobssonita e 2833 materiais secundários, utilizados para determinar os diagramas de fase e a estabilidade termodinâmica dos materiais alvo. Calculamos a estrutura eletrônica, determinando bandgap e massa efetiva dos portadores de carga. Verificou-se que, assim como para perovskitas, ao se alterar o halogênio da composição, controla-se o bandgap do material, sendo que este é maior para flúor, diminuindo na direção do iodo. Usando filtros relacionados a estabilidade e aplicação do material, selecionamos sete candidatos para futuras investigações: $MgAlCl_5$, $CaAlCl_5$,$SrAlCl_5$,$MgInF_5$,$CaInF_5$,$BeInF_5$ e $MgGaF_5$. Este estudo constitui um exemplo de busca em largura, com a utilização de experimentos *in silico* para explorar um vasto espaço de materiais. Essa abordagem permitiu a criação de uma nova base de dados, além de demonstrar a utilização da combinatória química para geração de espaço de busca.

Já no Capítulo 6, evidenciamos o uso de aprendizado máquina para a geração de novos materiais inorgânicos cristalinos de composição binária. Para a geração dos compostos utilizamos substituições químicas. Para a predição das propriedades (estabilidade e bandgap), utilizamos *features* atômicas para treinar algoritmos de aprendizado raso. As composições químicas criadas foram exploradas pelos modelos preditores, selecionando composições e sistemas cristalinos que representavam materiais estáveis candidatos para a aplicação. Com isso pudemos explorar mais de 1 milhão de composições, com tempo de experimentação *in silico* muito menor do que aquele necessário pelas técnicas de DFT *high througtput* - horas de vez de meses. Discutimos questões de interpretabilidade via uma boa análise exploratória dos *datasets* de treino e uso de estatísticas de correlação. O melhor algoritmo para a predição encontrado por nossos testes foi o RandomForest, tanto para predição de estabilidade quanto se o material era semicondutor ou metal. Testes com regressores diretos

do bandgap foram limitados pela quantidade de dados de treino e o conjunto reduzido de *features* utilizadas. Este estudo de caso constituiu apenas uma prova de conceito do domínio dessas técnicas para geração de novos materiais, sendo necessárias outras etapas para sua real utilização. Neste exemplo pudemos aumentar drasticamente o espaço de busca, mostrando as vantagens de se utilizar modelos substitutos para uma busca em largura.

Por fim, no Capítulo 7 sumarizamos os principais conceitos aprendidos, sugerindo futuros tópicos de aprofundamento para o avanço da área.

Os dados computacionais desta tese estão disponível em `https://github.com/ferreirahs/ciencia_dados_materiais`.

2 Abordagem Convencional

A abordagem convencional em ciências sustenta-se na inter-relação entre os métodos empíricos e os métodos teóricos. Dos métodos empíricos, a observação visa a catalogação do fenômeno, e a experimentação, sua reprodução. É com essas informações que podemos inferir as variáveis que o causam ou que se relacionam com ele. Os métodos teóricos, por sua vez, procedem por vias dedutivas a partir de regras gerais, estas podendo ser assumidas *a priori* ou estabelecidas através da experiência. Mais do que um conjunto de regras, o arcabouço teórico deve prover explicações racionais e, sobretudo, capacidade de predição. É da possibilidade de predição do comportamento de um fenômeno baseado em suas variáveis descritas pela teoria, que esta atinge um grau de maturidade e aceitação.

Uma característica importante dos métodos teóricos em Ciência de Materiais é sua formulação matemática, oriunda da física. Em geral, modelos que podem ser expressos por equações favorecem as deduções, possibilitando descrições e predições precisas dos fenômenos. Por exemplo, os modelos teóricos da Mecânica Quântica alicerçam o entendimento de átomos e moléculas, constituintes fundamentais dos materiais ordinariamente conhecidos, além de permitir a descrição e previsão de fenômenos como a radioatividade e a supercondutividade [34].

Com o advento do computador, uma nova possibilidade de se fazer ciência somou-se ao método empírico e o método teórico, um híbrido simbolizado pela simulação computacional. Mais fortemente ligado aos métodos teóricos, a simulação visa implementar o ferramental matemático destes modelos em algoritmos que possam ser executados pelo computador. Essa execução carrega, inclusive, um sentido de experimentação, sendo comum o uso do termo "experimentos computacionais". É desta etapa de experimentação *in silico* que diferentes materiais podem ser simulados e estudados.

O grau de precisão de uma simulação é dependente do modelo teórico assumido e das aproximações consideradas em sua implementação. Muitos esforços na área de Simulação Computacional de Materiais são feitos no sentido de tornar os métodos de simulação mais acurados com os valores obtidos na realidade. Por outro lado, estes métodos tem a vantagem de encurtar a distância entre os modelos teóricos e os dados experimentais, permitindo predições e descrições desses resultados. Por isso, a abordagem convencional em Ciência de Materiais é a utilização mista dessas três abordagens, experimento, teoria e simulação.

Neste Capítulo apresentamos uma revisão de métodos computacionais e suas teorias associadas, assim como técnicas de síntese e caracterização de materiais. O objetivo é condensar as informações a cerca dos principais métodos que foram usados para os estudos de caso discutidos neste trabalho. Além disso, visamos demonstrar um panorama que permitisse o contraste com as técnicas de Ciência de Dados de Materiais que são discutidas no Capítulo 3.

2.1 Simulação Computacional de Materiais

Simulação é um processo pelo qual um modelo matemático evolui de acordo com um conjunto de valores para determinadas variáveis. Por exemplo, podemos estar interessados em simular o comportamento de uma partícula puntual movendo-se livremente; para tanto, bastaria utilizar a expressão do movimento retilíneo uniforme (o modelo), avaliado em diferentes valores de tempo (a evolução), e obtendo assim toda sua trajetória sobre pontos passados e futuros; a execução da simulação, nesse sentido, expressou-se no cálculo das equações do modelo para diferentes situações de entrada.

Em geral, gostaríamos de simular um processo sobre o qual temos um modelo fidedigno à situação real. Entretanto, nem sempre temos disponível tal modelo, de forma que estamos lidando com uma construção abstrata cujas predições estão num limite de aproximação dos efeitos reais que pretendemos modelar.

A etapa de modelagem do problema que se pretende simular vê-se associada ao domínio de conhecimento do problema específico. Se pretendemos simular um fenômeno econômico, é preferível que o modelo proposto esteja ancorado em conhecimentos dessa ciência. Para materiais, portanto, os modelos são construídos sobre a Física e a Química, uma vez que essas ciências estabelecem as leis gerais por trás dos constituintes da matéria.

No que se convencionou chamar de Ciência Computacional de Materiais os modelos matemáticos se fundamentam em alguns arcabouços teóricos que variam de acordo com a escala do sistema que está sendo simulado. Simulações de objetos macroscópicos como sólidos estão no limite do contínuo, se valendo de equacionamentos próprios como aqueles associados aos métodos de Elementos Finitos (do inglês, *Finite Element Method*, FEM) [35, 36]. Em tais métodos, os fenômenos físicos podem ser descritos por equações diferenciais clássicas, sejam aquelas referentes a resistência mecânica dos materiais, a transferência de calor, a dinâmica dos fluídos[1] ou a propagação eletromagnética. Por outro lado, quando pretende-se simular um sistema em escala menor, na resolução de moléculas e átomos, os fenômenos físicos precisam ser modelados por métodos semi-clássicos e, principalmente, quânticos. Como exemplos, podemos citar a Teoria do Funcional da Densidade, a Dinâmica Molecular, e os Métodos de Hartree-Fock.

Uma abordagem integralista com objetivos de engenharia exigiria que, para um design completo de um material para uma determinada aplicação, todos os níveis de resolução pudessem ser explorados. Dessa forma, as simulações realizadas poderiam ter seus resultados integrados hierarquicamente para fornecer uma descrição precisa do comportamento do material na situação alvo. Apesar dos profundos avanços vistos nas últimas décadas, ainda não é possível realizar essa abordagem como um todo, tendo sido realizado apenas para casos mais simples.

Tendo em vista essa limitação, abordaremos a seguir apenas os fundamentos dos métodos de simulação com maior aderência a nossa pesquisa. Uma vez que nossa motivação

[1] Para fluídos pode ser computacionalmente mais eficiente aplicar outro método clássico, o FVM (*Finite Volume Method*).

original envolvia a construção de LEDs, as propriedades que estamos mais interessados são aquelas referentes a emissão de luz por semicondutores. Considerando os mecanismos físicos envolvidos nesse processo, devemos tratar o problema por métodos quânticos a fim de resolver a estrutura eletrônica do material.

2.1.1 Estrutura Eletrônica

Um dos objetivos da Química Quântica é a obtenção de soluções da Equação de Schrödinger (Eq. 2.1) para a determinação de propriedades de sistema atômicos e moleculares [37]. Nesta equação, na qual $\vec{r}$ representa o vetor posição em um dado sistema de coordenadas e t um instante de tempo, o comportamento de uma partícula de massa m submetida a um potencial $V(\vec{r},t)$, é determinado por sua função de onda $\Psi(\vec{r},t)$ através de[2]:

$$-\frac{\hbar^2}{2m}\nabla^2\Psi(\vec{r},t) + V(\vec{r},t)\Psi(\vec{r},t) = i\hbar\frac{\partial}{\partial t}\Psi(\vec{r},t). \tag{2.1}$$

Em uma abordagem simplificada, não-relativística, a função de onda Ψ de uma única partícula com massa de repouso m e energia total E submetida a um potencial independente do tempo $V(\vec{r})$, pode ser separada como $\Psi(\vec{r},t) = \psi(\vec{r})\phi(t)$, de forma que a Equação de Schrödinger do sistema pode ser dividida em um sistema de duas equações diferenciais:

$$\Big[-\frac{\hbar^2}{2m}\nabla^2 + V(\vec{r})\Big]\psi(\vec{r}) = E\psi(\vec{r}), \tag{2.2}$$

$$i\hbar\frac{d}{dt}\phi(t) = E\phi(t), \tag{2.3}$$

onde a Eq. 2.2 é o estado estacionário da parte espacial e a Eq. 2.3 é a evolução da fase ϕ ao longo do tempo.

Deste resultado, podemos expressar de maneira geral soluções estacionárias não-relativísticas (Eq. 2.2) utilizando álgebra matricial, como:

$$\hat{H}\psi = E\psi, \tag{2.4}$$

onde $\hat{H} = \Big[-\frac{\hbar^2}{2m}\nabla^2 + V(\vec{r})\Big]$ é o operador hamiltoniano e E é o autovalor (correspondente a energia) do autovetor de estado ψ.

Para determinar a probabilidade de se encontrar a partícula em uma determinada região do espaço, é necessário integrar o produto da função de onda da partícula por seu conjugado, $|\Psi(\vec{r},t)|^2 = \Psi(\vec{r},t)\Psi^*(\vec{r},t)$, sobre todo o intervalo considerado. Por outro lado, para se obter o valor esperado do momento linear $\langle p\rangle$ de uma partícula, por exemplo, confinada em uma dimensão, é necessário aplicar o operador $-i\hbar\frac{\partial}{\partial x}$ sobre o estado antes de integrá-lo sobre todo o intervalo. Assim se sucede para todos os observáveis físicos. [38]

Em materiais o objetivo é estudar sistemas atômicos e moleculares sujeitos aos princípios quânticos descritos anteriormente. Neste caso devemos especificar a natureza do

[2] $\hbar = h/2\pi$ é por vezes chamada de constante de Dirac, sendo h a constante de Planck.

potencial ao qual a partícula de interesse está submetida. Assumindo uma condição estacionária e não-relativística, queremos descrever um sistema de N elétrons e M núcleos atômicos usando esse formalismo. Considerando apenas as interações eletrostáticas entre essas partículas e suas respectivas energias cinéticas, temos que o hamiltoniano do sistema é, em unidade atômicas[3]:

$$\hat{H} = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \frac{1}{2}\sum_{k=1}^{M}\frac{1}{M_k}\nabla_k^2 - \sum_{i=1}^{N}\sum_{k=1}^{M}\frac{Z_k}{r_{ik}} + \sum_{i=1}^{N}\sum_{j>1}^{N}\frac{1}{r_{ij}} + \sum_{k=1}^{M}\sum_{l>k}^{M}\frac{Z_k Z_l}{R_{kl}}. \tag{2.5}$$

Nesta equação, os dois primeiros somatórios representam as energias cinéticas de elétrons e núcleos, respectivamente; o terceiro somatório define a atração eletrostática entre um núcleo de carga total Z_k e um elétron separados por uma distância $r_{ik} = |\vec{r_i} - \vec{r_k}|$; já o quarto e quinto somatórios estabelecem a repulsão entre pares de elétrons separados de $r_{ij} = |\vec{r_i} - \vec{r_j}|$ e pares de núcleos separados de $R_{kl} = |\vec{R_k} - \vec{R_l}|$, respectivamente.

Este problema de muitos corpos não pode ser resolvido analiticamente, sendo necessárias aproximações mesmo para a obtenção de soluções numéricas.

A primeira grande aproximação é aquela definida por Max Born e Robert Oppenheimer em 1927 [39]. Na aproximação de Born-Oppenheimer, observa-se que as velocidades dos elétrons são muito maiores do que as velocidades dos núcleos devido ao fato de que as massas dos núcleos M_k são muito maiores do que as massas dos elétrons m_i. Por exemplo, para um único próton $M \approx 1800m$. Isso implica que do ponto de vista dos elétrons, os núcleos estão praticamente fixos. Dessa forma podemos separar o problema de estudo do sistema atômico em dois problemas distintos: a dinâmica dos núcleos e a dinâmica dos elétrons.

Se desconsiderarmos a energia cinética dos núcleos e percebermos que a repulsão entre eles contribui por uma constante, poderemos reescrever a parte eletrônica do sistema como:

$$\hat{H} = -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 - \sum_{i=1}^{N}\sum_{k=1}^{M}\frac{Z_k}{r_{ik}} + \sum_{i=1}^{N}\sum_{j>1}^{N}\frac{1}{r_{ij}}, \tag{2.6}$$

Neste sentido, dizer que estamos resolvendo a estrutura eletrônica de um sistema atômico ou molecular do qual um material é constituído significa dizer que estamos resolvendo a Eq. 2.4 quando o hamiltoniano é definido pela Eq. 2.6. Apesar de o problema continuar um problema de muitos corpos (i.e., muitos elétrons), podemos nos valer dessa aproximação para mover a posição dos núcleos em etapas distintas da simulação do sistema atômico.

Usando o Teorema de Hellmann-Feynman [40, 41], podemos calcular as forças entre os átomos assumindo suas posições fixas em um determinado passo da simulação. Por exemplo, para calcular a componente radial da força F_r, poderíamos escrever o teorema utilizando notação bra-ket de Dirac como[4]:

[3] Neste sistema de unidades, as distâncias são dadas em raios de Bohr $a_0 = \epsilon_0 h^2/\pi m_e e^2$ e a energia em hartree $E_h = \hbar^2/m_e a_0^2$ [37].

[4] Aqui estamos expressando a ideia geral do teorema. Para um formalismo detalhado dentro do conceito

$$F_r = -\frac{\partial E}{\partial r} = -\left\langle \psi \left| \frac{\partial \hat{H}}{\partial r} \right| \psi \right\rangle . \tag{2.7}$$

Quando tratamos de uma estrutura cristalina, ao considerar a posição fixa dos núcleos, estamos assumindo volume ou pressão fixos para a geometria da célula [40]. O objetivo de minimizar a força é exatamente estabelecer a geometria da célula que é mais estável nessas condições, ou seja, a estrutura cristalina com menor energia.

Todavia, sabemos que a estrutura depende das condições ambientes de pressão e temperatura. Apesar de ser possível incluir uma pressão arbitrariamente modificando-se as posições dos núcleos fixos na rede cristalina do material ou adicionando-se alguns graus de liberdade a volume constate, incluir efeitos de temperatura pode ser não trivial. No caso de considerar seus efeitos devemos também resolver a dinâmica dos núcleos antes de resolver a estrutura eletrônica, considerando que os núcleos se movem com uma determinada distribuição de velocidades (por exemplo, a distribuição de Maxwell-Boltzmann). Se os núcleos forem considerados apenas classicamente, podemos lançar mão de métodos de Dinâmica Molecular clássica (MD), onde o movimento de cada átomo é descrito pelo formalismo newtoniano da dinâmica, isto é, $\vec{F} = d\vec{p}/dt = m_i \frac{d^2 q_i}{dt^2} = -\frac{\partial V(q)}{\partial q_i}$, no qual $\vec{F}$ representa o vetor força resultante, $\vec{p}$ o vetor momento linear e q_i as coordenadas generalizadas de cada átomo. Calculando-se a estrutura eletrônica pelo formalismo quântico, temos uma abordagem denominada Dinâmica Molecular *ab initio* cuja sigla é AIMD (do inglês, *ab initio Molecular Dynamics*) [43]. Métodos de AIMD podem executar passos de MD para os núcleos seguido de um método quântico para a estrutura eletrônica, ou executar uma dinâmica de Car–Parrinello (CP) onde equações de movimento para os orbitais moleculares são propagadas simultaneamente com as equações de movimento clássicas para os núcleos usando um formalismo lagrangiano [44, 43, 45]. O primeiro caso também é conhecido como BOMD (Born-Oppenheimer *Molecular Dynamics*) e o segundo caso como CPMD (Car-Parrinello *Molecular Dynamics*) [45], sendo ambos formas de se realizar AIMD. Em ambos os casos os elétrons são considerados no estado fundamental após a minimização da energia eletrônica.

Na sequência discutimos alguns dos principais métodos quânticos para solução da estrutura eletrônica.

2.1.1.1 Método de Hartree-Fock

O método de Hartree-Fock (HF) assume que a função de onda estacionária da porção espacial da estrutura eletrônica pode ser expressa pelo produto das funções de onda de cada um dos elétrons que constituem o sistema, denominado de Produto de Hartree, $\psi(\vec{r}) \approx \psi_1(\vec{r})\psi_2(\vec{r})...\psi_N(\vec{r})$. Entretanto, como elétrons são férmions, eles são indistinguíveis. Nesse sentido, é dito que os elétrons seguem uma distribuição de Fermi-Dirac respeitando o Principio de Exclusão de Pauli. Este princípio de indistinguibilidade dos

de DFT sugerimos a referência [42].

férmions é determinado matematicamente pelo fato de que a função de onda resultante do sistema deve ser anti-simétrica perante uma troca das coordenadas de dois elétrons[37], que é o mesmo que explicitar a interação de troca. John Slater introduziu em 1929 uma forma conveniente de expressar isso através da expansão em funções de determinantes de matrizes $N \times N$ contendo todas as funções de onda spin-orbitais $\chi(\vec{r})$ dos elétrons individuais [46]. Uma boa aproximação deste método consiste em escolher de forma ótima[5] apenas um determinante [37] para representar a função de onda $\psi(\vec{r}_1, \ldots, \vec{r}_N)$ da estrutura eletrônica, como apresentado na Eq. 2.8:

$$\psi(\vec{r}_1, \ldots, \vec{r}_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_1(\vec{r}_1) & \chi_2(\vec{r}_1) & \cdots & \chi_N(\vec{r}_1) \\ \chi_1(\vec{r}_2) & \chi_2(\vec{r}_2) & \cdots & \chi_N(\vec{r}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(\vec{r}_N) & \chi_2(\vec{r}_N) & \cdots & \chi_N(\vec{r}_N) \end{vmatrix}. \tag{2.8}$$

Considerando a função de onda da Eq. 2.8 e o hamiltoniano expresso na Eq. 2.6, queremos calcular:

$$E_{HF} = \frac{\left\langle \psi \middle| \hat{H} \middle| \psi \right\rangle}{\langle \psi | \psi \rangle}, \tag{2.9}$$

o que constitui encontrar o conjunto de orbitais χ_i que minimizam a energia esperada, ou seja, melhor aproximam o estado fundamental do sistema.

A energia cinética e de interação com o núcleo de um mesmo elétron i na Eq. 2.9 pode ser expressa por:

$$h_{ii} = \int \chi_i^*(\vec{r}) \left[-\frac{1}{2}\nabla_i^2 - \sum_{k=1}^{M} \frac{Z_k}{r_{ik}} \right] \chi_i(\vec{r})\, d\vec{r}. \tag{2.10}$$

Enquanto a integral de repulsão eletrônica entre dois elétrons i e j:

$$\langle \chi_i \chi_j | \chi_k \chi_l \rangle = \int \int \frac{\chi_i^*(\vec{r}_1)\chi_j(\vec{r}_1)\chi_k^*(\vec{r}_2)\chi_l(\vec{r}_2)}{|\vec{r}_1 - \vec{r}_2|}\, d\vec{r}_1 d\vec{r}_2. \tag{2.11}$$

A integral expressa na Eq. 2.11 pode ser dividida em dois termos: o termo de repulsão coulombiano, denotado por J, e o termo de troca, denotado por K:

$$J_{ij} = \int \int \frac{|\chi_i(\vec{r}_1)|^2 |\chi_j(\vec{r}_2)|^2}{|\vec{r}_1 - \vec{r}_2|}\, d\vec{r}_1 d\vec{r}_2, \tag{2.12}$$

$$K_{ij} = \int \int \frac{\chi_i^*(\vec{r}_1)\chi_j(\vec{r}_1)\chi_j^*(\vec{r}_2)\chi_i(\vec{r}_2)}{|\vec{r}_1 - \vec{r}_2|}\, d\vec{r}_1 d\vec{r}_2. \tag{2.13}$$

O termo repulsão J representa a repulsão eletrostática média entre as distribuições de

[5] Embora o Hartree-Fock escolha um único determinante ótimo, a solução exata da equação de Schrödinger muitas vezes exige uma combinação linear de vários determinantes de Slater. Essa ideia leva a métodos pós-Hartree-Fock, como a Interação de Configuração (CI) e a Teoria de Perturbação de Møller-Plesset (MP2), onde a função de onda é escrita como uma soma de múltiplos determinantes $\psi = c_0\psi_{\mathrm{HF}} + c_1\psi_1 + c_2\psi_2 + \ldots$. Um único determinante não é suficiente, por exemplo, em sistemas fortemente correlacionados ou para se calcular estados eletrônicos excitados.

carga dos elétrons nos orbitais i e j e não depende do spin dos elétrons. Já o termo de troca K surge da anti-simetria da função de onda do determinante de Slater e não tem equivalente clássico, sendo, portanto, um efeito puramente quântico. Este efeito faz os elétrons de mesmo spin evitarem interagir fortemente entre si, de maneira que o termo K só contribui se os orbitais χ_i e χ_j corresponderem a elétrons com o mesmo spin.

A energia total no método Hartree-Fock contém uma soma de termos de Coulomb e de troca:

$$E_{\mathrm{HF}} = \sum_i h_{ii} + \frac{1}{2}\sum_{i,j}(J_{ij} - K_{ij}). \tag{2.14}$$

O termo de Coulomb J_{ij} sempre aumenta a energia, pois representa a repulsão eletrônica. Já o termo de troca K_{ij} reduz a energia, criando um efeito de estabilização para elétrons de mesmo spin (por isso, orbitais de mesmo spin tendem a evitar sobreposição no espaço).

Não obstante, aplicar o método de Hartree-Fock, como explicado acima, traz uma complicação: a medida que o número N de elétrons aumenta, o custo computacional para se calcular a Eq. 2.14 cresce devido as integrais de quatro centros (Eq. 2.11). Logo a aplicação do método HF para sistemas com muitos elétrons como sólidos e proteínas é, normalmente, impraticável.

2.1.1.2 Teoria do Funcional da Densidade

Na Teoria do Funcional da Densidade (DFT), Pierre Hohenberg e Walter Kohn demonstraram, em 1964, que um sistema de muitos elétrons poderia ser descrito através de um funcional da densidade eletrônica do sistema [47]. Um ano depois dessa demonstração, hoje conhecida como Teorema de Hohenberg-Kown, em 1965, Kohn e Lu Jeu Sham desenvolveram um método autoconsistente para equacionar o problema e resolvê-lo através da Equação de Kohn-Sham [48]. Esse formalismo é necessário uma vez que a forma explicita do funcional exato da densidade é desconhecida [45], isto é, sabemos que o funcional exato existe, mas não conhecemos a sua forma explícita.

Na proposta de Kohn-Sham, primeiro devemos escrever a energia E_{KS} como um funcional da densidade eletrônica ρ, $E[\rho(\vec{r})]$:

$$E_{KS}[\rho] = T[\rho] + U[\rho] + \int V(\vec{r})\rho(\vec{r})d\vec{r}, \tag{2.15}$$

onde T denota a energia cinética dos elétrons, U a energia de interação elétron-elétron e V a energia do potencial eletrostático externo devido aos íons estáticos (núcleos).

Rescrevendo a expressão para separar os termos de energia cinética dos elétrons não interagente T_s e dos elétrons interagentes T_c, obtemos:

$$E_{KS}[\rho] = \big(T_s[\rho] + T_c[\rho]\big) + \big(U_H[\rho] + U_{xc}[\rho]\big) + \int V(\vec{r})\rho(\vec{r})d\vec{r}. \tag{2.16}$$

A Energia de Hartree U_H é o potencial eletrostático sentido por um elétron devido a todos os outros elétrons, assumindo que estes estejam fixos. Já $U_{xc} = U_c + U_x$ é o termo de correlação e troca, sendo que o termo de troca U_x é devido a interações entre elétrons através do Princípio de Exclusão de Pauli. Uma vez que elétrons são férmions, eles são indistinguíveis entre si, logo não há uma maneira de saber se um elétron trocou de lugar com outro elétron [49]. Da mesma forma, para respeitar o Princípio de Exclusão de Pauli, dois elétrons não podem apresentar todos os mesmos valores para seus números quânticos: para cada nível de energia, apenas dois elétrons são permitidos, um com spin para cima e outro com spin para baixo. A interação de troca mede a força que surge para manter dois elétrons de mesmo spin que estão em níveis de energia diferentes. Por outro lado, os termos de correlação U_c e T_c não são conhecidos em uma forma exata porque derivam das interações eletrônicas de muitos corpos[6]. Podemos evidenciar na equação os termos bem conhecidos (U_H and T_s) dos termos desconhecidos (U_x, U_c and T_c):

$$E_{KS}[\rho] = \big(T_s[\rho] + U_H[\rho]\big) + \big(T_c[\rho] + U_c[\rho] + U_x[\rho]\big) + \int V(\vec{r})\rho(\vec{r})d\vec{r}. \tag{2.17}$$

Substituindo o termo de energia de Hartree e juntando os termos de correlação e troca, $T_c[\rho] + U_c[\rho] + U_x[\rho] = E_{xc}[\rho]$:

$$E_{KS}[\rho] = T_s[\rho] + \int\int \frac{\rho(r)\rho(r')}{|r-r'|} d^3r d^3r' + \int V_{xc}(r)\rho(r)d^3r + \int V(r)\rho(r)d^3r. \tag{2.18}$$

sendo que o potencial de troca e correlação é determinado por $V_{xc}(\vec{r}) = \frac{\delta E_{xc}(\vec{r})}{\delta\rho(\vec{r})}$.

A densidade eletrônica ρ real do sistema deve ser aquela que minimiza a energia [50]. Dessa forma, aplicando o princípio variacional no funcional

$$\Omega[\rho] = E[\rho] - \mu\left(\int \rho(r)d^3r - N\right), \tag{2.19}$$

onde μ é o multiplicador de Lagrange, o problema torna-se encontrar uma densidade de carga ρ que satisfaça:

$$\int d^3r\Big(\Omega[\rho(r) + \delta\rho(r)] - \Omega[\rho(\vec{r})]\Big) = 0, \tag{2.20}$$

$$\int \delta\rho(r)\left(\frac{\delta T_s}{\delta\rho} + \int \frac{\rho(r')}{|r-r'|}d^3r' + V_{xc}[\rho]\right)d^3r = 0. \tag{2.21}$$

Pode-se escrever a densidade de carga ρ através dos orbitais de Kohn-Sham φ_i como apresentado na Eq. 2.22. Apesar de cada autovalor ε_i ser apenas o multiplicador de Lagrange, há um grande acordo entre os valores medidos experimentalmente e os valores

[6] A solução analítica completa da equação de Schrödinger para sistemas com múltiplos elétrons é intratável.

de energia preditos por ε_i. Em outras palavras, não há razão formal para interpretar o autoestados φ_i como autoestados reais dos elétrons, entretanto, a soma desses valores φ_i determina uma entidade real, a densidade eletrônica total:

$$\rho(\vec{r}) = \sum_{i=1}^{N} |\varphi_i(\vec{r})|^2 . \tag{2.22}$$

Dessa forma, o problema de muitos corpos inicial pode ser reduzido a um problema de um único corpo através da Equação de Kohn-Sham (KS):

$$\left(-\frac{1}{2}\nabla^2 + V_{KS}\right)\varphi_i(\vec{r}) = \varepsilon_i\varphi_i(\vec{r}), \tag{2.23}$$

onde φ_i e ε_i são os i-ésimo autovetor e autovalor da função de onda KS, respectivamente, e o potencial de Kohn-Sham é $V_{KS} = V_H(\vec{r}) + V_{xc}(\vec{r}) + V(\vec{r})$, como descrito nas passagens acima. Ou seja:

$$\hat{H}\varphi_i(\vec{r}) = \varepsilon_i\varphi_i(\vec{r}). \tag{2.24}$$

Deve-se notar que o método KS é aparentemente circular: é necessário saber a densidade eletrônica para se ter o potencial V_{KS}, e somente tendo tal potencial definido é que se torna possível calcular a função de onda KS (Eq. 2.23), que por sua vez define a densidade eletrônica (Eq.2.22). Logo, para resolver esse problema é gerada uma grade de densidade eletrônica $\rho_t(x, y, z)$ arbitrária. Para cada ponto no espaço, o valor de ρ pode ser atribuído usando-se números aleatórios ou orbitais atômicos como um chute inicial. O algoritmo de resolução tenta minimizar o erro entre o valor inicial e o valor calculado da densidade eletrônica pela Eq. 2.22, de maneira iterativa. Esse procedimento é conhecido como autoconsistente já que a densidade inicial deve ser igual a obtida como resultado após uma sequência de iterações, isto é, a medida que a densidade suposta ρ_t converge para o valor da densidade real ρ_0 da Eq.2.22.

O tipo de função usado para representar φ_i pode ser separado em espacialmente localizados e espacialmente estendidos, servindo para *clusters* e moléculas ou materiais bulk, respectivamente. Por exemplo, para cristais, condições periódicas de contorno são assumidas e são usadas ondas planas deslocalizadas para representar a função de onda:

$$\varphi_i(\vec{r}) = \sum_{\vec{G}} c_{i,\vec{G}} e^{i(\vec{k}+\vec{G})\cdot\vec{r}}, \tag{2.25}$$

onde $\vec{k}$ é o vetor de onda no espaço recíproco, $\vec{G}$ são os vetores da rede recíproca e $c_{i,\vec{G}}$ são os coeficientes da expansão. Observa-se que a representação na rede reciproca é quem encerra a qualidade periódica do sistema.

Com isso, a solução da Eq. 2.24 pode ser reescrita no espaço das ondas planas como um problema de diagonalização de matrizes:

$$\sum_{\vec{G}'} H_{\vec{G},\vec{G}'} c_{i,\vec{G}'} = \varepsilon_i c_{i,\vec{G}}\cdot, \tag{2.26}$$

onde $H_{\vec{G},\vec{G}'}$ é o hamiltoniano projetado na base de onda planas [7].

A representação da densidade eletrônica através da expansão em ondas planas (também conhecida em inglês como método PW, de *plane waves*) pode ser custosa computacionalmente. Na prática, apenas um certo conjunto de ondas é usado como base de representação dos estados, através de um corte de energia dado por $\frac{1}{2}|\mathbf{k}+\mathbf{G}|^2 \leq E_{\text{cut}}$. O potencial calculado com essa representação é denominado de pseudopotencial. Os pseudopotenciais substituem os potenciais de Coulomb da interação elétron-núcleo e elétron-elétron [44], podendo também levar em consideração elétrons de caroço até uma determinada energia de corte.

Já em métodos localizados, podemos combinar orbitais atômicos $\phi_\mu(\vec{r})$ para expandir a função de onda. Um desses métodos é a combinação linear dos orbitais atômicos conhecida como LCAO (do inglês, *Linear Combination of Atomic Orbitals*), que pode ser escrita como:

$$\varphi_i(\vec{r}) = \sum_\mu c_{i\mu}\phi_\mu(\vec{r}), \tag{2.27}$$

onde $\phi_\mu(\vec{r}$ são orbitais atômicos, como gaussianos ou funções de Slater, e $c_{i\mu}$ são os coeficientes de expansão.

Uma vez que os orbitais atômicos podem ser não ortogonais, devemos escrever à matriz de sobreposição:

$$S_{\mu\nu} = \int \phi^*_\mu(\vec{r})\phi_\nu(\vec{r})d\vec{r}. \tag{2.28}$$

A equação de Kohn-Sham no espaço LCAO se torna um problema de autovalores generalizado:

$$\sum_\nu H_{\mu\nu}c_{i\nu} = \varepsilon_i \sum_\nu S_{\mu\nu}c_{i\nu}, \tag{2.29}$$

onde $H_{\mu\nu}$ é o hamiltoniano projetado no espaço LCAO[8].

O método LCAO é dito localizado pois os orbitais são centrados nos núcleos dos átomos, seguindo a ideia de que os elétrons tendem a estar localizados ao redor dos átomos. Ele apresenta vantagem de ser menos demandante computacionalmente do que métodos PW, pois usa menos funções base, mas pode pode ser difícil garantir uma base suficientemente completa. Isso torna os métodos LCAO naturalmente adequado para moléculas e sistemas com falta de periodicidade, como superfícies ou defeitos, porém, menos eficiente para descrever sistemas como metais ou sólidos cristalinos.

[7] $H_{\vec{G},\vec{G}'} = \int e^{-i(\vec{k}+\vec{G})\cdot\vec{r}}\,\hat{H}\,e^{i(\vec{k}+\vec{G}')\cdot\vec{r}}\,d\vec{r}.$

[8] $H_{\mu\nu} = \int \phi^*_\mu(\vec{r})\hat{H}\phi_\nu(\vec{r})d\vec{r}.$

Além disso, quando é calculado a energia de Hartree, cada elétron contribui com o potencial influenciando todos os outros elétrons, inclusive ele mesmo. Isto resulta em uma auto-interação artificial que modifica o resultado final do esperado. Esses erros de auto-interação causam deslocalização na densidade eletrônica que impactam na predição dos valores de energia de valência e condução para materiais bulk ou nos níveis HOMO (o orbital com maior energia ocupado por elétrons, do inglês, *Highest Occupied Molecular Orbital*) e LUMO (orbital com a menor energia desocupado, do inglês, *Lowest Unoccupied Molecular Orbital*) de moléculas. Um exemplo importante deste problema é o da determinação do gap de energia E_G entre as bandas eletrônicas de condução e valência em um sólido cristalino: o método DFT por PBE-GGA tende a subestimar essa propriedade, devido a a descontinuidade derivativa Δ_{xc}. Podemos escrever esse problema como:

$$E_G = E_g^{KS} + \Delta_{xc}, \tag{2.30}$$

onde o gap real E_G é definido pela diferença entre o potencial de ionização I e a eletroafinidade A, $E_G = I - A$, e o gap calculado pelo DFT E_g^{KS} é definido pela diferença entre os autovalores do mínimo da banda de condução ϵ_{CBM} e do máximo da banda de valência ϵ_{VBM}, $E_g^{KS} = \epsilon_{CBM} - \epsilon_{VBM}$ [51].

Uma vez que a forma exata do funcional de troca e correlação $E_{xc}[\rho]$ não é conhecida, é necessário ajustá-lo empiricamente através de resultados experimentais ou construí-lo a partir de sistemas modelo [37]. Há diversas propostas: na aproximação de densidade local (*Local-Density Approximation*, LDA, em inglês), a energia de troca E_x é baseada no modelo de Thomas-Fermi-Dirac para interações em um gás homogêneo de elétrons (*homogeneous electron gas*, HEG) [50, 52]. Sua fórmula é:

$$E_x^{LDA}[\rho] = \frac{3}{4}\left(\frac{3}{\pi}\right)^{\frac{1}{3}} \int \rho(r)^{\frac{4}{3}} d^3r. \tag{2.31}$$

O termo de correlação $E_c = T_c + U_c$ é calculado de maneira acurada através de métodos de Monte Carlo Quântico [52]. Para limites de baixa e alta densidade, o valores de E_c podem ser ajustados, respectivamente, por:

$$E_c^{LDA}[\rho] = A\log(r_s) + B + r_s(C\log(r_s) + D), \tag{2.32}$$

$$E_c^{LDA}[\rho] = \frac{1}{2}\left(\frac{g_0}{r_s} + \frac{g_1}{r_s^{3/2}} + ...\right), \tag{2.33}$$

onde A, B, C, D, g_0 e g_1 são constantes ajustadas numericamente, e r_s é o parâmetro de Wigner-Seitz relacionado a densidade eletrônica ρ através de $r_s = \sqrt[3]{\frac{3}{4\pi\rho}}$.

É possível fazer o o método LDA também levar em consideração o spin dos elétrons através de uma abordagem denominada LSDA (do inglês, *Local Spin-Density Approximation*)[37, 53], na qual a densidade de carga ρ é separada em duas porções que representação o spin para cima $\rho_\uparrow$ e o spin para baixo $\rho_\downarrow$.

Já que a grade simulada pode ter uma densidade de carga descontínua, o método de aproximação por gradiente generalizado (*Generalized Gradient Approximation*, GGA, em inglês) adiciona ao modelo LDA uma expressão para o gradiente da densidade de carga entre duas células da grade. Isto é, queremos adicionar informações sobre a não-homogeneidade da densidade eletrônica ao se calcular $\nabla\rho(\vec{r})$ [37]. O parâmetro s de não-homogeneidade pode ser expresso por:

$$s = \frac{|\nabla\rho(\vec{r})|}{(16\pi^2\rho^4)^{1/3}}. \tag{2.34}$$

O funcional de troca LDA será corrigido então para:

$$E_{xc}^{GGA}[\rho] = \frac{3}{4}\left(\frac{3}{\pi}\right)^{\frac{1}{3}} \int \rho(r)^{\frac{4}{3}} F(s) d^3r. \tag{2.35}$$

Um dos funcionais mais precisos é aquele devido a J. Perdew, K. Burke e M. Ernzerhof [53, 54], também chamado de PBE-GGA, no qual $F(s)$ é expresso por:

$$F(s) = 1 + \kappa - \frac{\kappa}{1 + \frac{\mu s^2}{\kappa}}, \tag{2.36}$$

onde s é dado pela Eq. 2.34, $\kappa = 0.804$ e $\mu \approx 0.21951$ [37].

De maneira geral, ao considerar o spin podemos expressar essa correção como:

$$E_{xc}^{GGA}[\rho_\uparrow, \rho_\downarrow] = \int \rho(r) V_{xc}(\rho_\uparrow, \rho_\downarrow, \nabla\rho_\uparrow, \nabla\rho_\downarrow) d^3r. \tag{2.37}$$

Vale destacar que, a natureza *ab initio* do método DFT depende do funcional de troca e correlação, cuja escolha definirá se o método DFT é puramente fundamentado na teoria que descreve seus equacionamentos ou em parâmetros semi-empíricos ajustados de dados experimentais. Notadamente, PBE é um funcional *ab initio* [37]. Ao longo da história, mais de 500 aproximações foram propostas sendo o PBE uma das mais aceitas [51]. Por exemplo, o PBEsol [55] (PBE para sólidos) é uma variação do funcional PBE construído especificamente para simulação de sólidos que tem melhor acurácia na predição de energias de coesão, parâmetros de rede e o módulo volumétrico, mas não gerando bons resultados para sistemas moleculares.

Outra maneira de melhorar as simulações DFT-GGA é incorporar no modelo mais informações sobre a densidade eletrônica, adicionando o laplaciano da densidade eletrônica ∇^2 e a densidade de energia cinética $\tau(r)$:

$$E_{xc}^{Meta-GGA}[\rho] = \int \rho(r) V_{xc}\big(\rho(r), \nabla\rho(r), \nabla^2\rho(r), \tau(r)\big) d^3r. \tag{2.38}$$

A Eq. 2.38 expressa forma geral de um funcional denominado meta-GGA. Exemplos de implementações deste funcional são o funcional LAP [37], o SCAN (do inglês, *Strongly Constrained and Appropriately Normed*) [56] e o r2SCAN (uma versão regularizada do SCAN para corrigir instabilidades numéricas) [57].

2.1.1.3 Métodos Híbridos

Além dos métodos meta-GGA, as soluções do DFT com funcionais PBE podem ser melhoradas utilizando-se funcionais híbridos. Este funcionais são uma combinação linear das energias obtidas por métodos PBE e HF. A fração de cada termo considerada na combinação é um peso denominado de parâmetro α. Por exemplo, podemos expressar a energia de correlação e troca como $E_{xc} = \alpha E_x^{HF} + (1-\alpha)E_x^{PBE} + E_c^{PBE}$. Existem algumas formas de realizar essas misturas, como os funcionais híbridos PBE0, o B3LYP e HSE06 [37, 51, 58]. O funcional HSE, por exemplo, tem seu nome devido aos proponentes, Heyd-Scuseria-Ernzehof, sendo expresso através de uma combinação das energias de troca e correlação da seguinte forma:

$$E_{xc}^{HSE} = \alpha E_x^{HF,SR}(\omega) + (1-\alpha)E_x^{PBE,SR}(\omega) + E_x^{PBE,LR} + E_C^{PBE}, \tag{2.39}$$

onde SR se refere a curto alcance (do inglês, *short-range*) e LR refere-se a longo alcance, (*long-range*).

Enquanto o HSE06 é mais adequado para materiais periódicos, utilizando um modelo de troca modificado com uma separação de distâncias curtas e longas, o método B3LYP é mais utilizado em sistemas moleculares. Neste método B refere-se ao funcional B88 desenvolvido por Arhtur Becke [59], LYP refere-se ao funcional desenvolvido por Lee, Yang e Parr [60] e o número 3 refere-se ao fato de que está sendo feita uma combinação de três termos principais (troca HF, troca B88 e correlação de LYP):

$$E_{xc}^{B3LYP} = \alpha E_x^{HF} + (1-\alpha)E_x^{B88} + E_C^{LYP}. \tag{2.40}$$

Um dos motivos para a utilização de métodos híbridos é que no método KS a auto-interação do elétron não é nula, enquanto no método HF essa interação é completamente cancelada pelo termo de troca [37]. Funcionais como B3LYP ou HSE, combinam a troca HF com a troca-correlação da DFT de forma que, enquanto a troca de Hartree-Fock reduz o erro da autointeração, já que funcionais puramente locais (LDA, GGA) superestimam a delocalização eletrônica (o que leva a gaps de energia menores do que os reais), a correlação DFT mantém eficiência computacional, já que a parte da correlação vinda de funcionais locais melhora a descrição de interações eletrônicas sem o alto custo da teoria de perturbação ou métodos de correlação explícita. Essa abordagem permite, portanto, obter valores de bandgap em maior concordância com os valores experimentais. Entretanto, a natureza semiempírica (*ad hoc*) dos funcionais híbridos, tanto em sua modelagem quanto na escolha do parâmetro de mistura, podem gerar discussões quanto a qualidade *ab initio* destes métodos. Em outras palavras, o fato da calibração dos parâmetros de mistura dos funcionais ser realizada empiricamente para se obter uma maior concordância com os valores experimentais pode ser visto como uma deterioração da qualidade de teoria baseada em primeiros princípios (a descrição física exata do sistema) - isto é, estamos modificando o parâmetro α para ajustar a previsão aos dados experimentais.

2.1.1.4 Teoria de Perturbação de Muitos Corpos

Uma outra aproximação para a auto-interação do elétron é a Aproximação GW (do inglês, *GW Approximation*), onde G representa a função de Green para o sistema de um corpo e o W representa a interação Coulombiana dinâmica de resposta linear (interação efetiva). Nesta aproximação, a auto-energia do elétron é um produto de G e W. Isso significa que a partícula adicional (elétron ou buraco) não percebe partículas individuais do sistema, mas uma densidade de carga, e somente a resposta linear dessa densidade de carga é considerada. Normalmente, calcula-se usando a aproximação da fase aleatória (do inglês, *Random Phase Approximation*, RPA), mas às vezes outras abordagens, como a Teoria Funcional da Densidade Dependente do Tempo, são usadas para determinar W. [61]

A utilização de função de Green $G(r, r'; E)$ é um recurso matemático a solução da equação diferencial que determina a evolução do sistema. Por exemplo, consideremos a Equação de Schrödinger de estado estacionário na forma[9]:

$$\big(H(r) - E\big)\psi(r) = f(r), \tag{2.41}$$

onde H é o hamiltoniano, E a energia, ψ a função de onda do sistema, supomos um problema cuja inomogeneidade $f(r)$ é definida como uma função Delta de Dirac:

$$\big(H(r) - E\big)G(r, r'; E) = \delta(r - r'). \tag{2.42}$$

Assim, se $f(r) = V(r)\psi(r)$, podemos escrever:

$$\psi(r) = \psi_0(r) + \int G(r, r'; E)V(r)\psi(r)dr'. \tag{2.43}$$

Recursivamente, podemos assumir $\psi(r)$ dentro da integral como $\psi_0(r)$. Repetindo a iteração indefinidamente, resultando na Equação de Dyson:

$$\psi(r) = \psi_0(r) + \boldsymbol{GV}\psi_0(r) + \boldsymbol{GVGV}\psi_0(r) + \boldsymbol{GVGVGV}\psi_0(r) + ..., \tag{2.44}$$

onde $\boldsymbol{G}$ e $\boldsymbol{V}$ são as representações matriciais.

Podemos escrever a Função de Green $G(r, r'; E)$ fazendo a expansão em autovetores conhecidos para outro problema, por exemplo $\big(H(r) - E\big)\psi(r) = 0$:

$$G(r, r'; E) = \sum_n \frac{\psi_n(r)\psi_n(r')}{E - E_n}. \tag{2.45}$$

Para um problema dependente do tempo, isto é, uma solução com função de onda não necessariamente estacionária:

[9] Aqui seguimos a formulação matemática exposta em [62] e [63].

$$\big(H(r) - i\hbar\frac{\partial}{\partial t}\big)G(r,t;r',t') = \delta(r-r')\delta(t-t'). \tag{2.46}$$

Assim como para o caso independente do tempo, a solução dessa equação diferencial pode ser escrita como:

$$\psi(r,t) = \int G(r,t;r',t')\psi(r',t')d^3r'. \tag{2.47}$$

Aqui temos expressa a evolução temporal, que leva um estado em (r,t) para o instante no tempo e espaço (r',t'). É dessa capacidade de descrever a transição entre estados do sistema que faz a função de Green ser conhecida como propagador. No caso, uma evolução temporal expressa por $U(t,t') = e^{-\frac{i}{\hbar}H(t-t')}$ poderia ser entendida por:

$$G(r,t;r',t') = \langle r|e^{-\frac{i}{\hbar}H(t-t')}|r'\rangle \tag{2.48}$$

Já a relação entre $G(r,r';E)$ e $G(r,t;r',t')$ é expressa pela transformada de Fourier:

$$G(r,t;r',t') = \frac{1}{2\pi}\int G(r,r';E)e^{-\frac{i}{\hbar}E(t-t')}dE. \tag{2.49}$$

Assim, a expansão na base de estados conhecida será:

$$G(r,r';E) = \sum_n \frac{\psi_n(r)\psi_n(r')}{E - E_n \pm i\varepsilon}, \tag{2.50}$$

onde $\pm i\varepsilon$ define se estamos tomando polos positivos (correspondendo a função adiantada de Green, G^A) ou negativos (correspondendo a função retardada de Green, G^R).

Como conclusão desta seção, podemos afirmar que a introdução da função de Green no contexto da DFT fornece uma ferramenta poderosa para tratar correlações eletrônicas além das limitações da aproximação de Kohn–Sham. A formulação GW permite incorporar efeitos de auto-energia de forma sistemática, descrevendo a interação do elétron com a densidade eletrônica do sistema e permitindo cálculos mais precisos de energias de excitação e gaps eletrônicos. Através das equações de Dyson e das representações dependentes e independentes do tempo, a função de Green atua como propagador, conectando estados estacionários e não estacionários, e possibilitando uma interpretação física clara da evolução do sistema. Dessa maneira, a combinação da DFT com métodos baseados em funções de Green constitui uma ponte essencial entre abordagens de funcional e técnicas de muitos corpos, ampliando significativamente a capacidade preditiva da teoria para sistemas complecos, como o tratamento de muitos corpos.

2.1.1.5 Teoria do Funcional da Densidade Dependente do Tempo

A Teoria do Funcional da Densidade Dependente do Tempo (do inglês, *Time-dependent Density Functional Theory*, TDDFT) é uma extensão da DFT que lida com sistemas quânticos fora do equilíbrio e com processos dinâmicos. Enquanto a DFT tradicional é usada para sistemas estacionários (ou seja, com estados de energia bem definidos), a TDDFT é

usada para estudar a resposta temporal de sistemas quânticos a estímulos externos, como campos elétricos, luz ou outras perturbações.

A TDDFT baseia-se no teorema de Runge–Gross, que estabelece que, para um dado estado inicial, a densidade eletrônica dependente do tempo $\rho(\vec{r}, t)$ determina unicamente o potencial externo $V(\vec{r}, t)$ que gera essa densidade, até uma constante dependente do tempo [64, 65]. Isso permite descrever a evolução temporal do sistema utilizando equações de Kohn–Sham dependentes do tempo, semelhantes às equações estacionárias da DFT, porém incorporando explicitamente a dependência temporal. Na prática, a TDDFT é amplamente utilizada para calcular espectros de excitação, respostas ópticas, forças induzidas por campos externos e propriedades de transporte eletrônico em sistemas moleculares e sólidos. Além disso, quando combinada com aproximações de função de troca-correlação adequadas, a TDDFT fornece uma maneira de estudar fenômenos dinâmicos de muitos corpos sem a necessidade de tratar explicitamente todos os estados excitados do sistema.

Embora a TDDFT ofereça um formalismo elegante e relativamente eficiente para descrever a dinâmica de sistemas quânticos e calcular propriedades excitadas, sua aplicação prática a sistemas com muitos átomos ainda apresenta desafios [66]. A principal limitação decorre do custo computacional associado à solução das equações de Kohn–Sham dependentes do tempo, que cresce rapidamente com o número de elétrons no sistema. Além disso, a precisão dos resultados depende criticamente da escolha do funcional de troca-correlação dependente do tempo, que nem sempre captura corretamente efeitos de correlação de longo alcance ou excítons fortemente ligados. Em sistemas grandes, essas limitações podem levar a aproximações adicionais, truncamentos de bases ou simplificações que reduzem a fidelidade da descrição dinâmica, tornando o método adequado principalmente para moléculas de tamanho moderado ou supercélulas reduzidas de sólidos.

2.1.1.6 Softwares para Métodos Quânticos

Para realizar as simulações dos materiais é necessário ter, a principio, uma estrutura que quantifica posições espaciais de uma coleção de espécies químicas iguais ou diferentes. Em estruturas cristalinas, a periodicidade permite estabelecer um número reduzido de parâmetros suficientes para a reprodução de toda a estrutura. Os parâmetros de rede são definidos pelos comprimentos das arestas da célula unitária (a, b, c) e pelos ângulos entre as arestas (α, β, γ). Os vetores da base são definidos em coordenadas cartesianas como $\vec{a} = a\hat{x}$, $\vec{b} = b\hat{y}$ e $\vec{c} = c\hat{z}$ para um sistema ortogonal. Para sistemas não ortogonais temos:

$$\vec{a} = a\hat{x}, \tag{2.51}$$

$$\vec{b} = b\big(\cos\gamma\hat{x} + \sin\gamma\hat{y}\big), \tag{2.52}$$

$$\vec{c} = c\left(\cos\beta\hat{z} + \mu\hat{y} + \sqrt{1 - \cos^2\beta - \mu^2}\hat{z}\right), \tag{2.53}$$

onde $\mu = \frac{\cos\alpha - \cos\beta\cos\gamma}{\sin\gamma}$.

Já as coordenadas do i-ésimo átomo são designadas por $\vec{r_i} = x_i\hat{x} + y_i\hat{y} + z_i\hat{z}$ ou, em coordenadas fracionarias, como $\vec{r_i} = u_i\vec{a} + v_i\vec{b} + w_i\vec{c}$.

Estes elementos geométricos são utilizados para especificar, normalmente em um arquivo à parte, a estrutura do material a ser simulado. Além disso, neste arquivo, são inseridos caracteres (convencionalmente o símbolo químico) que definem o átomo cujas coordenadas estão sendo especificadas. Isso é importante para que durante um cálculo DFT, por exemplo, seja possível recuperar o respectivo pseudopotencial atribuído a cada átomo[10].

Para implementar as equações físicas que regem o comportamento dos elétrons e átomos conforme apresentado nas seções anteriores, existem alguns softwares comerciais e livres. Na Tabela 2.1 apresentamos alguns dos principais:

Tabela 2.1: Softwares para simulações quânticas separados por licença (adaptado de [19]).

Comercial	GPL†	MIT ‡
VASP	Quantum Espresso	CPMD
CASTEP	ABINIT	GAMESS
ONETEP	CP2K	FLEUR
Gaussian	BigDFT	
Molpro	SIESTA	
Turbomole	octopus	
CRYSTAL	GPAW	
Q-Chem	exciting	
FHI-aims		
WIEN2k		
ORCA		

† *GNU General Public License.* ‡MIT License.

2.1.2 Dinâmica Molecular

Dinâmica Molecular (MD) é uma técnica computacional utilizada para simular o comportamento de sistemas de partículas, como átomos e moléculas, ao longo do tempo, com base nas leis da mecânica clássica. Nesse método, as posições e velocidades das partículas são atualizadas iterativamente por meio da integração numérica das equações de movimento de Newton, considerando as forças derivadas de potenciais interatômicos ou campos de força previamente definidos [67]. Esses potenciais modelam interações como ligações covalentes, forças de van der Waals, interações eletrostáticas e efeitos de solvatação.

[10]Um mesmo átomo pode ter mais de um pseudopotencial, cuja distinção está nas aproximação consideradas durante a construção do pseudopotencial em si. Por exemplo, pode-se considerar o potencial de um átomo de Cl usando elétrons de última camada ou adicionando efeitos de elétrons de caroço. Isso exige que seja usado um corte (*cutoff*) de energia maior na representação de ondas planas.

Matematicamente, as equações de trajetória podem ser obtidas das Equações de Newton:

$$m_i \frac{d^2 q_i}{dt^2} = -\frac{\partial V(q)}{\partial q_i}, \tag{2.54}$$

em que q_i são as coordenadas generalizadas, m_i a massa e $V(q)$ o potencial sentido pelo átomo i.

Também é possível empregar o formalismo hamiltoniano:

$$\begin{aligned} \frac{dp_i}{dt} &= -\frac{\partial H}{\partial q_i}, \\ \frac{dq_i}{dt} &= \frac{\partial H}{\partial p_i} \end{aligned} \tag{2.55}$$

em que q_i são as coordenadas generalizadas, p_i o momento linear e H o hamiltoniano.

Na dinâmica molecular, as equações de movimento — como as equações de Hamilton ou de Newton — são integradas numericamente para obter a trajetória das partículas ao longo do tempo. Como essas equações geralmente não admitem soluções analíticas para sistemas com muitos corpos, utilizam-se métodos de integração numérica, especialmente os que preservam propriedades físicas do sistema, como energia e momento.

Um dos métodos mais usados é o algoritmo de Verlet (e suas variantes, como o *Velocity Verlet*), que fornece uma boa conservação de energia e é simples de implementar[11]. A ideia central é atualizar iterativamente as posições e velocidades das partículas com base nas forças que atuam sobre elas. O algoritmo de integração transforma essas derivadas temporais em diferenças finitas para passos discretos de tempo Δt. Abaixo está um exemplo, considerando a posição q_i, momento p_i, massa m, e força $F_i = -\frac{\partial H}{\partial q_i}$:

$$\begin{aligned} q_i(t+\Delta t) &= q_i(t) + \frac{p_i(t)}{m}\Delta t + \frac{1}{2m} F_i(t)\Delta t^2 \\ F_i(t+\Delta t) &= -\frac{\partial H}{\partial q_i}(q_i(t+\Delta t)) \\ p_i(t+\Delta t) &= p_i(t) + \frac{1}{2}\left[F_i(t) + F_i(t+\Delta t)\right]\Delta t \end{aligned} \tag{2.56}$$

No entanto, além de simplesmente resolver essas equações, é fundamental que a simulação represente corretamente o comportamento estatístico do sistema em condições termodinâmicas específicas. É nesse contexto que entram os ensembles, que definem quais variáveis macroscópicas permanecem constantes durante a simulação. Por exemplo, ao utilizar o algoritmo de integração Velocity Verlet para atualizar posições e velocidades, pode-se combinar esse esquema com mecanismos adicionais — como termostatos e barostatos — que servem como reservatório para equilíbrio de temperatura e pressão, respectivamente [69]. As velocidade e posições iniciais são também estabelecidas como distribuições estatísticas específicas (como a distribuição de Maxwell-Boltzmann para a velocidade de átomos ou moléculas em gases ideais). Dessa forma, é possível realizar simulações que não apenas obedeçam à mecânica clássica, mas também amostrem corretamente o espaço de fases

[11] Este é o algorítimo implementado no software de simulação de dinâmica molecular LAMMPS [68].

compatível com as condições termodinâmicas desejadas.

Utiliza-se, portanto, o conceito de ensemble para descrever os sistemas que serão simulados por MD. Na mecânica estatística, um ensemble é uma representação teórica de um grande número de cópias virtuais de um sistema físico, cada uma em um microestado diferente, mas todas compatíveis com certas condições macroscópicas. Os ensembles são classificados de acordo com as variáveis termodinâmicas que permanecem constantes: no ensemble microcanônico (NVE), o número de partículas N, o volume V e a energia E são fixos; no canônico (NVT), a energia pode flutuar enquanto a temperatura T é mantida constante; e no ensemble isotérmico-isobárico (NPT), tanto o volume quanto a energia flutuam, mantendo constantes N, P (pressão) e T.

Além da escolha do ensemble, a fidelidade de uma simulação de dinâmica molecular depende diretamente do campo de força (do inglês, *force field*, FF) utilizado, que define como as interações entre os átomos são modeladas. Campos de força clássicos, como AMBER, CHARMM, OPLS e COMPASS, parametrizam ligações covalentes, ângulos, torções, bem como interações de van der Waals e eletrostáticas, ajustando-se a dados experimentais ou resultados de cálculos quânticos [70]. A escolha do potencial é crucial, pois determina a precisão das propriedades estruturais, dinâmicas e termodinâmicas obtidas.

Outro aspecto importante é a escala temporal e espacial acessível. Tipicamente, simulações de MD clássica utilizam passos de tempo da ordem de femtossegundos (10^{-15} s), permitindo atingir escalas de tempo de nanossegundos a microssegundos, dependendo dos recursos computacionais disponíveis. Isso é suficiente para estudar difusão[71], relaxações locais, transições de fase estrutural [72] e processos conformacionais em biomoléculas [73], mas insuficiente para fenômenos mais lentos, como enovelamento completo de proteínas ou difusão iônica em sólidos. Para contornar essas limitações, são empregados métodos acelerados, técnicas de amostragem aprimorada ou mesmo acoplamento com modelos *coarse-grained*[12][74].

Embora a MD clássica seja extremamente eficiente, ela ignora explicitamente efeitos quânticos, como a formação e quebra de ligações químicas. Para tratar sistemas nos quais as interações eletrônica desempenha papel essencial, utiliza-se a Dinâmica Molecular *ab initio* (como discutido na Seção 2.1.1), na qual as forças são obtidas diretamente de cálculos de estrutura eletrônica, geralmente via DFT [43]. Uma alternativa intermediária é o método Car–Parrinello, que unifica a dinâmica eletrônica e nuclear em um mesmo formalismo [44], ou ainda abordagens híbridas QM/MM (*quantum mechanics/molecular mechanics*), que tratam regiões críticas com mecânica quântica e o restante do sistema com potenciais clássicos.

Finalmente, vale destacar que simulações realistas de MD envolvem um grande número de partículas, frequentemente superiores a centenas de átomos, o que demanda técnicas de paralelização massiva e o uso de supercomputadores ou GPUs. O desenvolvimento de algoritmos otimizados, como decomposição espacial ou múltiplas escalas de tempo, tem

[12]Nestes modelos, as moléculas são compostas não por átomos individuais, mas por grupos de átomos tomando como um único "pseudo-átomo".

sido essencial para ampliar a aplicabilidade da MD a sistemas complexos de interesse em ciência dos materiais e biologia molecular.

2.2 Estabilidade Termodinâmica de Materiais

Para determinar se um material pode existir, seja de ocorrência natural ou artificial, é necessário verificar a sua estabilidade. Entretanto, o termo estabilidade pode expressar uma grande variedades de ideias. Mais precisamente, a estabilidade termodinâmica refere-se ao equilíbrio termodinâmico entre o material e o ambiente que o circunda [75, 76]. Para nossos propósitos, o equilíbrio termodinâmico é o mínimo global de energia para a estrutura iônica e eletrônica de um sólido cristalino, quando a temperatura ambiente é $T = 0$ K e a pressão absoluta $P = 0$ Pa. Apesar de o valor absoluto dessa energia quando obtida por um cálculo DFT estar associado a um referencial arbitrário, as diferenças de energias entre duas estruturas tem exatamente o significado físico que estamos procurando. Dessa forma, podemos definir o mínimo global de energia em relação as diferenças entre diferentes estruturas. Do ponto de vista químico, a formação de um composto, por exemplo, um composto binário $\mathrm{A}_{1-x}\mathrm{B}_x$, depende da presença de precursores unários A e B. Portanto, ao calcular a energia de uma estrutura unária A e de uma estrutura unária B, podemos comparar tais valores com a energia da estrutura binária $\mathrm{A}_{1-x}\mathrm{B}_x$ de tal forma que, se a energia do composto for menor do que das fases elementares, então o composto é mais estável do ponto de vista energético. Contudo, há ainda uma segunda forma de comparar essas energias, agora não mais em relação as fases elementares e sim com relação a outro composto binário oriundo dos mesmos precursores, por exemplo $\mathrm{A}_{1-y}\mathrm{B}_y$. Tratamos as diferenças entre fases elementares e compostos como *entalpia de formação* ou *energia de formação* [77], enquanto que as diferenças entre dois compostos com os mesmos elementos e estequiometrias ou estruturas diferentes, de *convex hull* - também conhecido como *Maxwell convex hull* [78].

É importante deixar claro que a entalpia de formação só considera os estados finais do processo de formação, de maneira similar a Lei de Hess para reações químicas. Além disso, se a fase elementar de referência for considerada como um único átomo, então temos um delta de energia diferente conhecido como *energia coesiva* [79]. Em outras palavras, a energia de formação mede a diferença de energia entre o material e suas estruturas precursoras elementares, enquanto a energia coesiva mede a diferença de energia para atomizar esse material. Ambas, contudo, não consideram etapas intermediárias desses respectivos processos.

2.2.1 Entalpia de Formação

A entalpia de formação $\Delta H_f(A_1, ..., A_L)$ pode ser entendida como a diferença de energia entre um material e seus L elementos constituintes. Quando $\Delta H_f < 0$ a formação do composto é favorável em vez de suas fases elementares. Matematicamente, podemos

expressar isso como:

$$\Delta H_f(A_{1,n_1}...A_{L,n_L}) = E_{total}(A_{1,n_1}...A_{L,n_L}) - \sum_i^L n_i E_{total}(A_i), \tag{2.57}$$

onde $E_{total}(A_{1,n_1}...A_{L,n_L})$ e $E_{total}(A_i)$ é a energia total por átomo do composto e da fase elementar, respectivamente, e n é o número de átomos na estrutura do elemento químico A.

Por exemplo, para um material binário A_xB_y com uma estequiometria arbitrária x e y, nós temos:

$$\Delta H_f(A_xB_y) = E_{total}(A_xB_y) - \big(xE_{total}(A) + yE_{total}(B)\big), \tag{2.58}$$

onde $A_1 = A$, $A_2 = B$, $n_1 = x$ e $n_2 = y$ com respeito as variáveis da Eq. 2.57.

A entalpia de formação depende do ambiente químico definido pelas fases elementares precursoras. Por exemplo, podemos ter um mesmo precursor na forma molecular (em fase gasosa) ou na forma metálica (sólido puro). Na Tabela são apresentadas as fases que utilizamos em nossos cálculos.

Para ligas intermetálicas, os erros de precisão no valor E_{total} dos compostos e nas fases elementares de referência se anulam, fazendo com que os resultados computacionais correspondam aos experimentais [80]. Nestes casos, elementos e compostos estão ambos na forma metálica. Porém, quando os compostos e os elementos possuem características químicas diferentes, como um isolante iônico feito de metais e gases diatômicos, o cancelamento mútuo de erros não pode ocorrer [80, 81]. Por isso, neste trabalho utilizamos a correção FERE da fase elementar proposta por Stevanović et al. [80]. Essa correção é dada por:

$$\mu_i^{FERE} = \mu_i^{GGA+U} + \delta\mu_i^{FERE}, \tag{2.59}$$

onde $\delta\mu_i^{FERE}$ é o valor tabelado para o átomo i reportado em [80] e μ_i^{GGA+U} é a energia total da respectiva fase unária i calculada com DFT.

2.2.2 Convex Hull

Uma entalpia de formação negativa é um critério necessário, mas não suficiente, para determinar a estabilidade de um material: compostos com outras estequiometrias e/ou estruturas podem ter uma energia de formação mais baixa. Estes outros compostos são referidos como fases concorrentes do material alvo. Um composto preferirá se decompor em suas fases concorrentes quando elas tiverem energia de formação menor que a do composto.

A superfície convexa que liga todos os compostos com a energia mais baixa para uma estequiometria específica é chamada de casco convexo (do inglês, *convex hull*). Isto significa que os materiais acima do casco convexo são instáveis, enquanto os materiais no casco convexo são considerados estáveis.

Por outro lado, usando este critério, alguns materiais que realmente existem não estão no mínimo local numa referência de energia do *convex hull*. De acordo com a literatura atual, esses materiais são conhecidos como fases *metaestáveis*. Para nossos propósitos, definimos um material metaestável como uma estrutura até 100 meV acima do casco, seguindo *Sun et al.* [79]. No entanto, esta abordagem pode ser melhorada conforme descrito por *Aykol et al.* [82] usando o *limite amorfo*. Este limite deve ser calculado para cada uma das estequiometrias pretendidas para determinar o liminar de energia no qual as estruturas ainda podem ser metaestáveis, i.e., estruturas com energia menor do que a energia do limite amorfo podem existir (estáveis ou metaestáveis).

2.2.3 Estabilidade Dinâmica

A estabilidade dinâmica é um critério distinto, relacionado ao comportamento das vibrações atômicas. Na estrutura molecular ou cristalina, os átomos ocupam posições de equilíbrio com relação as forças de interação interatômicas. Entretanto, para uma temperatura finita, a posição real dos átomos não é estática: ela está oscilando em torno do ponto de equilíbrio. Essa oscilação é quantizada em quasipartículas denominadas fônons [34]. Dessa forma, a estabilidade dinâmica de um material pode ser analisada por meio da dispersão de fônons que existem neles (modos de vibração permitidos).

Em geral, as oscilações são consideradas harmônicas, ou seja, a variação da posição de equilíbrio é modelada por um oscilador harmônico. Contudo, em sistemas reais, é comum existirem anarmonicidades relacionadas a forma do potencial de interação.

Do ponto de vista teórico, se o espectro de fônons calculado por simulação computacional contiver frequências imaginárias (frequentemente representadas como valores negativos em gráficos), isso indica uma tendência da estrutura atômica a se distorcer espontaneamente, sinalizando uma fase dinamicamente instável. Isso não significa necessariamente que o material não possa existir — pode, em vez disso, sugerir uma transição de fase para uma configuração mais estável, especialmente em temperaturas mais baixas ou sob diferentes pressões. Além disso, frequências negativas podem surgir quando a estrutura do material não foi relaxada adequadamente, por erros no critério de convergência, por artefatos de interpolação, em sistemas onde os elétrons estão fortemente correlacionados, ou quando há acoplamento spin-fônon [83].

A obtenção do espectro vibracional e dispersão de fônons do material pode ser por via de simulações de Dinâmica Molecular [84], de Métodos Quânticos como DFT [85], ou de métodos AIMD [86].

2.2.4 Outras Estabilidades

Além dos critérios de estabilidade discutidos anteriormente, um material pode ser estudado, por exemplo, de acordo com sua estabilidade mecânica, térmica, química e elétrica. Nestes casos, considera-se um material macroscópico sujeito a um regime de estresse referente a qualidade de um destes fenômenos. Neste sentido, a estabilidade mecânica refere-se

à capacidade do material de resistir a forças externas sem sofrer deformações permanentes ou falhas estruturais; a estabilidade térmica está relacionada à capacidade do material de manter suas propriedades dentro de uma faixa de temperatura específica, evitando degradação ou mudanças indesejadas; a estabilidade química envolve a resistência do material a reações com substâncias externas, como corrosão ou oxidação; enquanto a estabilidade elétrica trata da capacidade do material de manter suas propriedades elétricas sob a ação de campos elétricos ou correntes elétricas elevados.

É preciso destacar que essas propriedades dependerão apenas da estrutura atômica e eletrônica quando o material é considerado em sua forma cristalina pura. Em geral, porém, características mesoscópicas terão um papel importante nessas propriedades. Por exemplo, em um material policristalino, os defeitos de interface terão um papel importante nas propriedades mecânicas, elétricas, térmicas e químicas.

Neste trabalho não exploramos essas outras qualidades de estabilidade, de forma que quando nos referimos a um material estável, estamos nos referindo a um material cuja entalpia de formação é negativa e que a sua distância de energia para o *convex hull* não é muito grande (i.e., em torno de 100 meV).

2.3 Síntese e Caraterização de Materiais

Materiais termodinamicamente estáveis ou metaestáveis podem ocorrer naturalmente nas condições de temperatura e pressão propícias. Dominar a capacidade de reproduzir esse ambiente abre a possibilidade de sintetizar esses mesmos materiais. Além disso, uma série de outros materiais são puramente sintéticos, oriundos do domínio de rotas de síntese em experimentos de laboratório e em escala industrial. A síntese de materiais é, portanto, uma etapa fundamental na ciência e engenharia de materiais, influenciando diretamente as propriedades estruturais, químicas e funcionais do produto final. Uma ampla variedade de técnicas de síntese foi desenvolvida, cada uma adaptada a classes específicas de materiais e morfologias-alvo.

A caraterização, por sua vez, é a etapa de determinação das propriedades físicas, químicas e estruturais dos materiais, sejam eles puramente sintéticos ou de ocorrência natural. Existem diferentes técnicas de caracterização, algumas mais simples e diretas e outras de alta sofisticação.

Numa abordagem convencional, um determinamo material ou composto químico é obtido por uma rota de síntese e em seguida caracterizado de acordo com um conjunto de técnicas. Esse processo pode ser repetido diversas vezes até que o material alvo seja sintetizado e as propriedades desejadas sejam alcançadas.

Esta seção tem como objetivo fazer uma breve revisão de algumas técnicas de síntese e caracterização sem a pretensão de ser completa e exaustiva. Algumas das técnicas descritas aqui foram usadas no estudo de caso I (Capítulo 4), no qual sintetizamos perovskitas híbridas pseudo-2D do tipo Dion-Jacbson.

2.3.1 Rotas de Síntese

As sínteses solvotérmica e hidrotérmica são abordagens amplamente utilizadas baseadas em solução aquosa, normalmente envolvendo a reação de precursores em um solvente em temperaturas elevadas e pressões autógenas dentro de um recipiente selado (como um autoclave). Na síntese hidrotérmica, a água atua como solvente, permitindo a cristalização de materiais como zeólitas e óxidos metálicos. Por outro lado, os métodos solvotérmicos expandem essa abordagem utilizando uma gama mais ampla de solventes orgânicos ou inorgânicos, facilitando a síntese de materiais complexos, incluindo estruturas metal-orgânicas (do inglês, *Metal-organic Frameworks*, MOFs) e calcogenetos [87].

Em contextos menos complexos, métodos de precipitação em solução em recipiente aberto podem ser usados para sintetizar certos compostos inorgânicos em temperaturas moderadas e sob pressão ambiente. Essa abordagem envolve a dissolução de precursores em um solvente e a aplicação de energia térmica por meio de uma placa de aquecimento, com agitação. Ao atingir a temperatura desejada, a precipitação do composto-alvo pode ocorrer quase instantaneamente devido a mudanças na solubilidade ou na cinética da reação. Embora não ofereça a precisão de métodos selados ou sob alta pressão, essa técnica proporciona simplicidade e processamento rápido.

Uma estratégia de cristalização baseada em solução que tem sido cada vez mais utilizada é a cristalização por temperatura inversa (do inglês *Inverse Temperature Crystallization*, ITC). Esse método se aproveita de solutos cuja solubilidade diminui com o aumento da temperatura, em contraste com o comportamento convencional. Materiais como perovskitas de haletos de chumbo (por exemplo, $MAPbBr_3$) têm sido sintetizados utilizando ITC a partir de solventes como $(CH_3)_2NC(O)H$ (dimetilformamida, DMF) ou $(CH_3)_2SO$ (dimetilsulfóxido, DMSO), onde o aquecimento induz a nucleação e o crescimento rápidos de cristais de alta qualidade [88].

Outra rota da química úmida é o processo sol–gel, que permite a síntese em temperaturas relativamente baixas a partir de precursores moleculares. Nessa abordagem, reações de hidrólise e policondensação (reação de polimerização) formam uma suspensão coloidal (sol) que evolui para uma rede gelatinosa (gel). A secagem subsequente e o tratamento térmico controlado produzem pós porosos, filmes finos ou cerâmicas monolíticas com homogeneidade composicional fina e estrutura nanométrica ajustável. O método sol–gel é amplamente utilizado na preparação de nanocompósitos. Além disso, sua capacidade de incorporar dopantes e híbridos orgânico-inorgânicos em nível molecular o torna uma rota atrativa para aplicações avançadas nas áreas óptica, eletrônica e de energia.[89, 90]

Diferente das rotas baseadas em solução, a síntese no estado sólido (ou solução sólida) envolve a reação direta de precursores sólidos — frequentemente óxidos, carbonatos ou nitratos — que são misturados intimamente (por meio de moagem, trituração ou ativação mecanoquímica) e, em seguida, submetidos a tratamentos térmicos em altas temperaturas. Reações impulsionadas por difusão durante a calcinação ou sinterização levam à formação da fase desejada. Embora normalmente exija temperaturas elevadas e tempos de reação

prolongados, a síntese no estado sólido oferece simplicidade para a preparação de materiais policristalinos como óxidos do tipo perovskita supercondutores e cátodos para baterias.

Já as técnicas de crescimento a partir do derretimento, como o método de Kyropoulos, a técnica de Bridgman–Stockbarger e o processo Czochralski, são usadas principalmente para a produção de cristais únicos em larga escala. O método de Kyropoulos envolve a extração lenta de um cristal-semente a partir do derretido sob gradientes de temperatura controlados, sendo frequentemente aplicado na produção de grandes cristais ópticos, como a safira (Al_2O_3) [91]. De maneira semelhante, a técnica de Bridgman–Stockbarger se baseia na solidificação direcional do derretido dentro de uma ampola selada que é movida através de um gradiente térmico controlado, permitindo o controle preciso da orientação e pureza do cristal [92]. O processo Czochralski consiste em puxar um cristal-semente de um banho fundido giratório, controlando simultaneamente a temperatura e a taxa de extração, o que possibilita a obtenção de cristais grandes e de alta qualidade, amplamente utilizados na fabricação de semicondutores, como silício e germânio, e semicondutores UWBG como o $\beta-Ga_2O_3$ [93].

Os métodos de fluxo utilizam um sal fundido ou fluxo de óxido como um solvente de alta temperatura no qual os precursores se dissolvem e cristalizam durante o resfriamento [94]. Essa abordagem é particularmente útil para o crescimento de cristais únicos de alta qualidade de óxidos ou compostos intermetálicos, pois o fluxo pode controlar a solubilidade e a cinética de nucleação da fase desejada.

A fusão zonal é outra técnica baseada no derretimento para purificação e crescimento de cristais, onde uma região estreita fundida é movida ao longo de um lingote sólido [95]. As impurezas se concentram na região fundida e são transportadas para uma das extremidades do lingote, permitindo o crescimento de cristais altamente puros e livres de defeitos — especialmente no processamento de semicondutores (como silício e germânio).

Além das rotas baseadas em solução, estado sólido e fusão, as técnicas de deposição em fase vapor são fundamentais na síntese de filmes finos e revestimentos nanoestruturados. A Deposição Química em Fase de Vapor (do inglês *Chemical Vapor Deposition*, CVD) envolve a reação química de precursores gasosos próximos ou sobre a superfície aquecida de um substrato, resultando na deposição de um filme sólido. Essa técnica é amplamente utilizada na indústria de semicondutores para a fabricação de materiais eletrônicos, optoeletrônicos e de energia, incluindo filmes de diamante, grafeno e dicalcogenetos de metais de transição bidimensionais [96]. Já a Deposição Física em Fase de Vapor (do inglês *Physical Vapor Deposition*, PVD), que inclui técnicas como evaporação e pulverização catódica (conhecida do inglês como *sputtering*), baseia-se na transferência física de material de uma fonte para o substrato em condições de vácuo. É comumente aplicada em microeletrônica, ótica e em revestimentos duros.

O crescimento epitaxial, especialmente por meio da Deposição Química de Vapor de Compostos Organometálicos (do inglês, *Metal-Organic Chemical Vapor Deposition*, MOCVD) ou da Epitaxia por Feixe Molecular (do inglês *Molecular Beam Epitaxy*, MBE), é uma técnica crucial para a fabricação de heteroestruturas semicondutoras com precisão

atômica [13]. Esses métodos possibilitam o crescimento de camadas com correspondência de rede, frequentemente exigidas em dispositivos de alto desempenho, como diodos emissores de luz (LEDs), lasers e transistores de alta mobilidade eletrônica.

2.3.2 Técnicas de Caracterização

A caracterização de materiais constitui etapa essencial para a compreensão de suas propriedades estruturais, morfológicas, químicas, ópticas, mecânicas, elétricas e térmicas. As curtas descrições de algumas dessas técnicas que é feita a seguir foram baseadas nas referências [14, 34, 97, 98].

A difração de raios-X (DRX) permite a identificação de fases cristalinas, parâmetros de rede, microdeformações e tensões internas. O raio-X incidente no material é refletido ao ser espalhado pela nuvem eletrônica dos átomos. As ondas refletidas podem estar em fase ou fora de fase de acordo com a camada atômica na qual ocorreu a reflexão, surgindo efeitos de sobreposição construtiva ou destrutiva responsáveis pelo padrão de difração. Matematicamente isso é expresso pela lei de Bragg, que descreve a relação para a qual ocorre corra a difração, sendo definida pelo o ângulo de difração θ da onda eletromagnética incidente de comprimento de onda λ e a distância d entre os planos atômicos do cristal:

$$2d\sin(\theta) = n\lambda, \tag{2.60}$$

onde n é um número natural não nulo que representa a ordem de difração.

Um espectro DRX corresponde a uma sequência de picos de intensidade de difração com relação ao ângulo θ. Esse espectro carrega a assinatura da estrutura cristalina do material, podendo ser refinado para a determinação do sistema cristalino e grupo espacial. Normalmente, as amostras caracterizadas por DRX são policristalinas (difração de pó) ou monocristalinas, porém existem variações e aperfeiçoamentos da técnica, por exemplo, permitindo difração de raio-X de superfícies.

Para a caracterização da morfologia pode-se empregar técnicas de microscopia ótica e microscopia eletrônica. A microscopia eletrônica de varredura (MEV) fornece imagens de alta resolução da superfície, possibilitando avaliar a morfologia, a distribuição de grãos e defeitos, além de análises químicas locais por espectroscopia de energia dispersiva (do inglês, *Energy Dispersive X-ray Spectroscopy*, EDS). Já a microscopia eletrônica de transmissão (MET) amplia a resolução para a escala atômica, permitindo investigar defeitos cristalinos e interfaces. Para ambas as técnicas, a preparação da amostra é um fator crítico: no caso do MEV, geralmente é necessário recobrir a superfície com uma fina camada condutora, enquanto para o MET são necessárias lamelas extremamente finas (da ordem de dezenas a centenas de nanômetros).

A microscopia de força atômica (do inglês, *Atomic Force Microscopy*, AFM) atua em escala nanométrica, mapeando a topografia de superfícies a partir da interação entre uma ponta sensível e a amostra. Existem vários modos de operação, desde a medida de força entre a ponta e a superfície do material por deslocamento ou por mudança da vibração da

ponta até medidas tunelamento quântico de corrente elétrica entre a ponta e o material quando este é condutor ou semicondutor (técnica está conhecida por microscopia de varredura por tunelamento, do inglês *Scanning Tunneling Microscopy*, STM). Essas técnicas são fundamentais para investigar rugosidade, adesão e propriedades mecânicas locais, sobretudo em materiais nanoestruturados. Para garantir a qualidade das medidas, as amostras devem assegurar superfícies os mais limpas, planas e estáveis possíveis de modo a minimizar interferências e grande saltos durante a varredura.

No estudo de sistemas coloidais, a determinação do potencial zeta fornece informações sobre as interações eletrostáticas entre partículas dispersas, sendo obtido a partir da mobilidade eletroforética. O potencial zeta é definido como o potencial elétrico no plano de cisalhamento da dupla camada iônica que envolve as partículas em suspensão. Esse parâmetro descreve a condição eletrostática da superfície dispersa em meio líquido e fornece uma medida indireta da carga efetiva das partículas. De forma complementar, a técnica de espalhamento dinâmico de luz (do inglês, *Dynamic Light Scattering*, DLS) permite obter a distribuição dos raios hidrodinâmicos, utilizada para estimar o tamanho das partículas e investigar processos de agregação em sistemas coloidais.

Entre as técnicas espectroscópicas, a absorção óptica na região ultravioleta-visível (UV-VIS) é utilizada para investigar bandas eletrônicas, estimar gaps ópticos e determinar concentrações de espécies químicas em solução. Essa técnica baseia-se na absorção de radiação eletromagnética por elétrons em átomos ou moléculas, promovendo transições eletrônicas entre diferentes níveis de energia, possibilitando a estimativa do gap óptico por meio da análise da borda de absorção. Em sistemas em solução, a absorção segue a lei de Beer–Lambert, que relaciona a absorbância à concentração da espécie química e ao caminho óptico, permitindo quantificação precisa de analitos. A medida é realizada em espectrofotômetros, nos quais um feixe de luz monocromática atravessa a amostra e a intensidade transmitida é comparada à de uma referência. Normalmente a referência é dita "branco", sendo medida a transmitância somente com o recipiente vazio no caminho óptico - isto é, mede-se a absorção de luz do recipiente e do ar circundante.

Já a espectroscopia de fotoluminescência (do inglês, *photoluminescence*, PL) consiste na excitação óptica de um material, geralmente semicondutores ou nanocristais, por meio de fótons de energia superior ao seu gap eletrônico. Os portadores fotoexcitados relaxam para estados de menor energia e recombinam radiativamente, emitindo fótons cuja energia corresponde às transições eletrônicas do sistema. Dessa forma, a técnica fornece informações diretas sobre a estrutura de bandas, defeitos eletrônicos e processos de recombinação.

A recombinação em si, é um processo de emissão espontânea, que é determinado pelo coeficiente de transição de Einstein A. Assim a taxa de emissão de N fóton em um material excitado pode ser descrita como:

$$\left(\frac{dN}{dt}\right)_{radiativo} = -AN. \tag{2.61}$$

Levando em consideração processos radiativos e não radiativos a taxa total é:

$$\left(\frac{dN}{dt}\right)_{total} = -\frac{N}{\tau_R} - \frac{N}{\tau_{NR}}, \tag{2.62}$$

onde $\tau_R = \frac{1}{A}$ é o tempo de vida radiativo da transição eletrônica, e τ_{NR} é o tempo de vida não-radiativo, quando o elétron relaxa do nível de energia excitado sem emitir fótons - por exemplo, trocando energia com fônons da rede cristalina.

A eficiência de luminescência η_R do material (também chamada de eficiência quântica) é definida pela razão entre a a taxa de emissão radiativa e a taxa de emissão total, de forma que:

$$\eta_R = \frac{AN}{N\left(\frac{1}{\tau_R} + \frac{1}{\tau_{NR}}\right)} = \frac{1}{1 + \frac{\tau_R}{\tau_{NR}}}. \tag{2.63}$$

A luminescência eficiente requer que o tempo de vida radiativo seja muito menor do que o tempo de vida não radiativo, que pode ser entendido pela máxima quantidade de luz emitida quando $\tau_R \ll \tau_{NR}$.

Isso é importante para a caraterização de materiais que serão aplicados em LEDs ou células solares, pois estabelece um limite para a eficiência que poderá ser atingida pelos dispositivos construídos com esse material.

Neste caso, para determinar o tempo de vida radiativo é necessária utilizar fotoluminescência resolvida no tempo (do inglês, *Time-Resolved Photoluminescence*, TRPL). Nesta técnica, a amostra é excitada por pulsos ultracurtos de laser e o espectro de emissão é registrado como função do tempo. Isso possibilita investigar a dinâmica de relaxação e recombinação de portadores, permitindo determinar tempos de vida radiativos e não radiativos com resolução temporal que pode alcançar a escala de picossegundos.

Enquanto na técnica de PL tradicional, a excitação é mantida num comprimento de onda fixo (fóton de mesma energia), na espectroscopia de excitação por fotoluminescência (PLE), utiliza-se uma fonte de laser sintonizável para excitar a amostra em diferentes comprimentos de onda, registrando-se a intensidade da luminescência em uma energia fixa de emissão. Como o sinal é proporcional ao coeficiente de absorção no comprimento de onda do laser, a PLE permite obter espectros de absorção óptica mesmo em amostras onde medições diretas de transmissão não são viáveis, como em filmes finos sobre substratos opacos, sendo portanto, uma técnica complementar a absorção por UV-VIS.

No que se refere ao transporte elétrico, medidas de resistividade, condutividade e mobilidade de portadores de carga são indispensáveis para compreender o comportamento de semicondutores. Essas medidas podem envolver uma vasta gama de técnicas. Em especial, um arranjo de quatro pontas metálicas pode ser montado sobre a amostra: duas pontas funcionam para injetar e coletar uma corrente elétrica através de uma fonte, enquanto as outras duas pontas são ligadas a um voltímetro sensível. Mantendo-se uma corrente fixa, a amostra pode ser submetida a diferentes temperaturas, de forma que a resistência elétrica dada pela razão entre a corrente injetada e a tensão medida, é determinada para cada

um desses pontos. Gráficos de resistência em função da temperatura demonstram assinaturas tipicas: retas crescentes com a temperatura configuram comportamento de metais, enquanto decrescimento exponencial caracteriza semicondutores. Dependendo da características intrínseca ou extrínseca, que pode ser determinada pelas regiões de temperatura, é possível calcular a energia de ativação e o bandgap desses semicondutores.

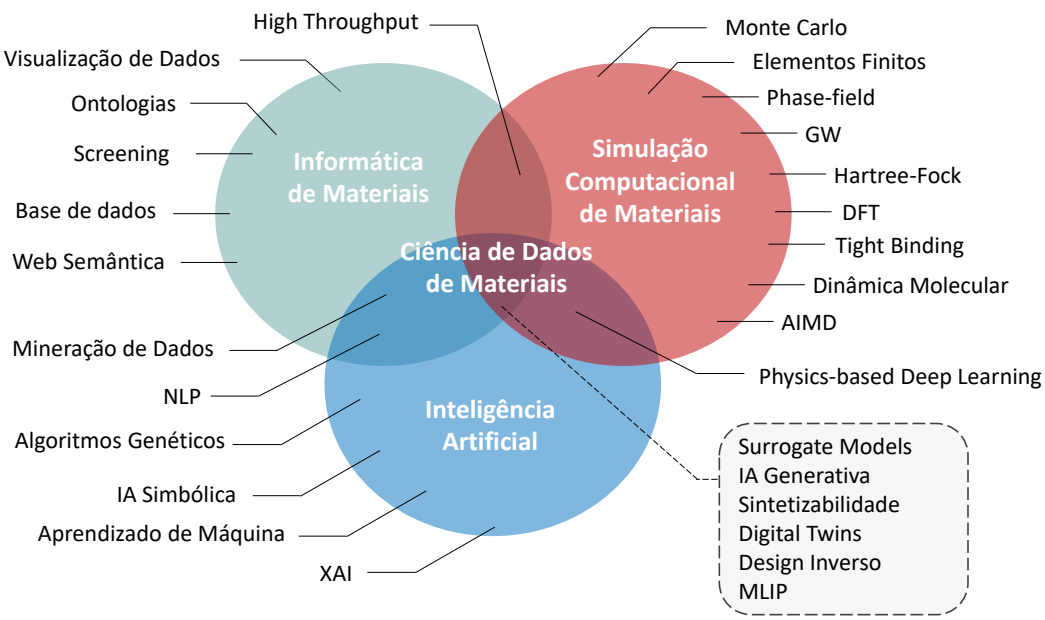
Outras medidas elétricas podem envolver quatro pontas, porém agora a uma temperatura fixa e um campo magnético aplicado perpendicular ao sentido da corrente, com a tensão sendo medida nos terminais ortogonais a linha de injeção. Neste caso, o Efeito Hall é utilizado para determinar o tipo de portador de carga majoritário (se elétrons ou buracos), além da concentração de portadores e sua mobilidade.

Finalmente, as técnicas termoanalíticas, em particular a termogravimetria (TG), são aplicadas para avaliar a estabilidade térmica, a decomposição e o teor de voláteis. A TG mede a variação de massa da amostra em função da temperatura. Durante a análise, perdas de massa podem ocorrer devido à liberação de gases adsorvidos ou absorvidos, à desidratação ou à degradação química dos constituintes do material. Outra técnica relevante é a diferencial de variação de calor (do inglês *Differential Scanning Calorimetry*, DSC), que mede fluxos de calor associados a transições físicas ou químicas, como fusão, cristalização ou reações exotérmicas e endotérmicas, complementando a avaliação da estabilidade e comportamento térmico dos materiais.

3 Ciência de Dados de Materiais

Ciência de Dados de Materiais (MDS do inglês, *Materials Data Science*) é como denominamos a sobreposição de conhecimentos oriundos da Informática de Materiais, da Simulação Computacional de Materiais e da Inteligência Artificial, conforme esquematizado na Figura 3.1:

Figura 3.1: Sobreposição de conhecimentos de diferentes áreas para a definição de Ciência de Dados de Materiais.



Fonte: elaborado pelo autor.

A Informática de Materiais (MI, do inglês, *Materials Informatics*) visa estabelecer métodos de tecnologia da informação, notadamente aqueles usados para a criação e a manutenção de grandes volumes de dados, visando construir um ecossistema que acelere a descoberta de novos materiais. Essa catalogação informatizada é tanto sobre materiais estudados experimentalmente, quanto sobre materiais estudados computacionalmente.

Com o crescimento dessas bases de dados e das técnicas para divulgação e compartilhamento, novos horizontes despontaram com a possibilidade de utilização de métodos computacionais tradicionalmente atrelados à Ciência da Computação. Destacam-se aqueles relacionados à Mineração de Dados, à Inteligência Artificial, IA, (do inglês, *Artificial Intelligence*, AI) e ao Aprendizado de Máquina (do inglês, *Machine Learning*, ML). Métodos de Aprendizado de Máquina, por exemplo, são ferramentas computacionais de aprendizado

estatístico, que conseguem extrair padrões em dados multidimensionais. Uma vez que o volume de dados associado à Ciência de Materiais cresceu com o paradigma da Informática de Materiais, tais técnicas tornaram-se promissoras para análise preditiva e descritiva de propriedades de materiais. Por outro lado, técnicas desenvolvidas no âmbito da IA Simbólica permitem representar conhecimento em ontologias que podem ser usadas tanto na orientação das bases de dados, quanto auxiliar na interpretação e explicabilidade dos resultados de modelos de ML.

Devemos lembrar, todavia, que a Figura 3.1 é uma esquematização incompleta, primeiro por destacar apenas alguns dos pontos relacionados a cada área, e depois, mais importante, por não destacar outras áreas que contribuem e podem receber contribuições da Ciência de Dados de Materiais. Dois exemplos são: (1) a área de Ciência de Materiais Experimental que contribui com dados empíricos para compor bases de dados (Informática de Materiais), contribui e recebe contribuições diretas dos modelos teóricos por trás das simulações (Materiais Computacionais) e pode se valer de modelos de Inteligência Artificial para predição e descrição de resultados empíricos; e (2) a área de Computação Quântica que estabelece novos métodos de simulações de Materiais Computacionais e usa e desenvolve técnicas de Inteligência Artificial para algoritmos quânticos de aprendizado de máquina [99].

Ademais de sistematizar a área de Ciência de Dados de Materiais quanto aos seus constituintes fundamentais, como discutido acima, é proveitoso trazer alguns conceitos atrelados à Ciência de Dados como disciplina própria:

- Quanto à **geração de dados**, é interessante notar que estes podem ser produzidos por humanos ou produzidos por máquinas [100]. Este conceito está fortemente atrelado ao conceito de Internet das Coisas (IoT, *Internet of Things*) para coleta automática de dados em diferentes objetos e situações. No contexto da MDS, temos como exemplo a transição entre os resultados de simulações de materiais realizadas por humanos para aquelas realizados majoritariamente por máquinas através de scripts *high throughput*; apesar do grau de automação variar entre os trabalhos, há uma tendência clara de priorizar uma "linha de produção industrial de cálculos". Há também tratativas de se realizar o mesmo com dados experimentais [101, 102]. Boa parte dos dados, contudo, ainda são gerados por humanos operando máquinas (instrumentos de caracterização ou de síntese de materiais, computadores para a simulação de materiais). Automatizar essas tarefas pode abrir espaço para as etapas criativas do processo de investigação científica.

- Quanto à **complexidade da tarefa analítica**, é usual separar, em complexidade crescente, em análise descritiva (o que aconteceu), análise diagnóstica (por que aconteceu), análise preditiva (o que pode acontecer) e análise prescritiva (como fazer acontecer) [100]. No contexto de MDS, temos como exemplo o estudo da sintetizabilidade de um material, que parte de dados estabelecidos por experimentos ou simulações para determinar padrões de estabilidade (análise descritiva), utiliza-se de conceitos físico-químicos para explicar os padrões observados (análise diagnóstica),

pode usar simulações ou modelos de IA para predizer estabilidade de um novo material intencionado (análise preditiva) e, finalmente, pretende propor um protocolo de síntese para este material (análise prescritiva). Ou seja, em última instância, a automatização das tarefas também pode substituir a etapa criativa ao responder essa série de perguntas.

Se considerarmos a contraposição entre a geração de dados e a complexidade analítica, fica clara a presença de um lacuna de síntese, isto é, uma diferença entre os dados obtidos, a informação armazenada, o conhecimento aprendido e as aplicações geradas. O paradigma de ciência orientada a dados visa diminuir esse lacuna crescente ao longo do tempo [19].

Uma vez que os fundamentos da Simulação Computacional de Materiais, em especial aqueles pertinentes a este trabalho, foram apresentados no Capítulo 2, iremos discutir a seguir os fundamentos da Inteligência Artificial e da Informática de Materiais, finalizando este capítulo com uma visão geral dos principais conceitos que surgem da intersecção dessas três áreas.

3.1 Fundamentos de Inteligência Artificial

A definição de Inteligência Artificial é ampla, uma vez que o conceito de inteligência em si, é ainda tópico de discussão científica e filosófica. É comum os autores se esquivarem dessa definição, utilizando uma definição circular, por exemplo:

> "Inteligência Artificial é o estudo dos conceitos que permitem aos computadores serem inteligentes" – Patrick Henry Winston. [103]

> "Inteligência Artificial é o estudo de como fazer os computadores realizarem tarefas em que, no momento, as pessoas são melhores" – Elaine Rich. [104]

> "Inteligência Artificial é simplesmente a transferência das características da inteligência humana para as máquinas" – Robert Levine, Diane Drang e Barry Edelson. [105]

> "O objetivo [da AI] é construir máquinas que, ao resolver problemas, pareçam pensar" – João de Fernandes Teixeira. [106]

> "Inteligência Artificial é o estudo dos sistemas que agem de um modo que a um observador qualquer pareceria ser inteligente" – Ben Coppin. [107]

Essa circularidade já havia sido posta inicialmente por Alan Turing em 1950, quando, ao tentar responder a questão filosófica do que seria uma "máquina pensar", ele estabelece uma tarefa, hoje conhecida como Teste de Turing, como um critério para essa definição [108]: com a devida resolução dessa tarefa, um sistema artificial seria dito dotado de inteligência. Ou seja, a propriedade de inteligência do sistema é verificada de maneira

indireta com a resolução de uma tarefa que pressupõem-se necessitar de inteligência para ser resolvida. O estudo dessas tarefas e de outras definições de inteligência artificial está além do escopo deste trabalho, nos limitando aqui apenas a introdução deste tema. Vale, contudo, ressaltar que autores como Stuart Russel e Peter Norvig abrangem todas essas definições em um esquema de classificação, atribuindo às definições seus caráteres quanto a sistemas que pensam como humanos, sistemas que agem como humanos, sistemas que pensam racionalmente e sistemas que agem racionalmente [109]. Por exemplo, para estes autores a definição de Turing se enquadra em um sistema que age como humano. É evidente que para cada grupo de definição, há subáreas de pesquisa que resultam em artefatos (códigos ou máquinas) e conceitos que podem ser classificados de acordo com esse esquema. Dessa forma, ainda para Russel e Norvig, o grande campo da AI:

> "[...] preocupa-se não apenas com a compreensão, mas também com a construção de entidades inteligentes - máquinas que podem calcular como agir de forma eficaz e segura em uma ampla variedade de situações inéditas." – Stuart Russel e Peter Norvig. [109]

Outro ponto importante nos fundamentos desta área de pesquisa é a *hipótese do funcionalismo* no qual a materialidade específica que suporta uma determinada capacidade (a inteligência) não é relevante, podendo esta ser realizada através de outros sistemas materiais [110]. É neste sentido que um "cérebro eletrônico" tem capacidade de apresentar inteligência (artificial) assim como um "cérebro biológico". Esta hipótese pode ainda ser reescrita como a computabilidade da inteligência, isto é, se inteligência pode ser recriada através de uma Máquina de Turing[1]. Apesar dessa restruturação da hipótese fornecer um formalismo matemático específico, continua em aberto a pergunta devido à imprecisão da própria definição de inteligência, como discutido anteriormente. Também é argumentado que há tarefas que não podem ser computadas, como o Problema de Parada[2] e que isto impediria a recriação de inteligência, porém isto só é verdade se de fato a inteligência for da mesma classe que o Problema de Parada [107]. Todavia, mesmo que um artefato apresente comportamento inteligente, pode-se argumentar em razão da compreensão, pelo próprio artefato, do processo como um todo. Esta objeção é denominada de Argumento do Quarto Chinês e foi proposta por John Searle em 1980, tendo diversas réplicas ao longo dos anos e estando fortemente atrelada à área da filosofia da mente e ao problema de qualia [106]. Diferentes aspectos dessa discussão são divididos numa dicotomia de IA forte e IA fraca: o primeiro sistema seria capaz de realizar diferentes tarefas (propósito geral) e seria indistinguível de um ser consciente enquanto o segundo sistema seria capaz de realizar somente tarefas específicas, sem demonstrar consciência. A hipótese da IA forte é definida pela capacidade de uma IA de fato ter uma mente e uma consciência enquanto a hipótese da

[1] O formalismo da Máquina de Turing estabelece que um computador pode ser implementado independentemente do substrato físico que realiza a computação de fato.

[2] Uma definição informal deste problema é: para qualquer par algoritmo e entrada finita, decida se aquele algoritmo terminará de executar ou se rodará indefinidamente.

IA fraca diz que, no máximo, uma IA irá parecer ter uma mente e uma consciência, sem de fato tê-las. Um sistema de IA forte hipotético é também denominado de AGI (*Artificial General Intelligence*).

Para a Ciência de Dados de Materiais vamos nos abster das discussões filosóficas e considerar, de maneira geral, que Inteligência Artificial é uma coleção de métodos computacionais que permitem, automática ou semi-automaticamente, realizar analises em diferentes tipos de dados, visando, em geral, sínteses descritivas, diagnósticas, preditivas ou prescritivas. A princípio, poderia ser dito que essa definição encerra-se no processamento de uma variedade de dados, não diferenciando-a, portanto, das ferramentas tradicionais da informática. Há contudo um pressuposto importante sobre o objeto da automatização: a substituição de tarefas cada vez mais complexas. É preciso lembrar, também, que essa coleção de métodos vem sendo acumulada e modificada ao longo dos anos, abrangendo diferentes categorias, de acordo com os pressupostos e intenções por trás da sua criação. Não só o avanço do poder computacional, porém também das abstrações desses métodos, permitiram que diferentes técnicas antes ditas de IA tornarem-se corriqueiras e ditas somente de computação ou informática (por exemplo, o paradigma de programação orientada a objetos que tem forte ligação com o conceito de quadros desenvolvido dentro do campo da IA [107]). Podemos ainda considerar, em uma simplificação, que inteligência artificial é um exemplo de conjunto de métodos de matemática aplicada, uma vez que tarefas automáticas realizadas por computadores são processos algorítmicos, isto é, passos matemáticos bem definidos, apesar dessa definição só destacar a importância dos conceitos matemáticos por trás destes métodos, sem de fato ajudar na sistematização dos conceitos propriamente desenvolvidos dentro da área de AI. O que destacamos, portanto, é que a IA provê artefatos para a MDS com caráter da automatização de tarefas complexas através da matemática aplicada, sem de fato estarmos preocupados em atestar inteligência a estes. Parafraseando Winston [103], estamos, no momento, mais interessados em "tornar os computadores mais úteis"do que de fato "entender os princípios que tornam a inteligência possível".

A fim de estabelecer uma visão geral sobre esses conjuntos de métodos de AI, iremos introduzir, a seguir, alguns dos conceitos ligados a duas categorias principais, destacando os pontos de interesses de cada um deles para a MDS.

3.1.1 Inteligência Artificial Simbólica

A inteligência artificial simbólica carateriza a primeira onda de desenvolvimento da área. Uma das hipóteses centrais nessa fase de desenvolvimento é a *hipótese de símbolos físicos*, na qual um sistema de símbolos físicos possui meios necessários e suficientes para a ação inteligente geral [104]. Do ponto de vista prático, a representação do símbolo e das expressões simbólicas no computador torna-se fundamental. Isso envolve tanto a estrutura de dados[3] quanto os conceitos da representação da informação (ou conhecimento) a que se destina o sistema de símbolos. Boa parte dos esforços dessa primeira onda estão atrelados

[3]Como estrutura de dados nos referimos a forma como computadores armazenam e organizam dados na memória [111], volátil ou permanente.

a representação do conhecimento[4], por exemplo, através de redes semânticas, quadros, a lógica proposicional, a lógica de predicados e a lógica não-monotônica [104, 107]. Também abarcam essas representações equações algébricas, equações diferenciais, resultados de simulações, invariâncias espaciais, relações de probabilidade e *feedback* humano [112]. Estes substratos simbólicos pode ser manipulados de acordo com regras específicas, cujas etapas são passiveis de serem automatizadas, o que permite construir sistemas que expressam raciocínio e planejamento.

Um dos primeiros sistemas a se utilizar destas técnicas foi o Logic Theorist desenvolvido em 1956 por Allen Newell, Herbert A. Simon e Cliff Shaw e que consistia em um programa capaz de encontrar provas para teoremas matemáticos [109]. Em 1964 temos o MATHLAB, o primeiro sistema CAS[5] desenvolvido em LISP[6] por Carl Engelman, que viria a desenvolver o Macsyma em 1968 e cujos exemplos atuais são Mathematica, Maple e Axiom.

Na área de materiais, temos como exemplo o DENDRAL, uma IA desenvolvida em 1969 por Ed Feigenbaum e usada para identificar moléculas orgânicas a partir de espectroscopia de massa [114]. Sistemas especialistas como o DENDRAL, são sistemas de conhecimento intensivo nos quais as soluções são obtidas através de um extenso conjunto de regras de propósito geral [109], devidamente representadas no computador e que podem ser combinadas para gerar as respostas.

3.1.1.1 Problemas de Busca

Um dos principais conceitos por trás da IA Simbólica é abstrair o problema original para um paradigma de solução de problema diferente, por exemplo, traduzindo o problema original em um problema de busca[103, 109]. Neste tipo de abordagem, é assumido um estado objetivo dentro de um espaço de estados finito; estes estados podem ser representados como V vértices (ou nós) de um grafo $G(V, A)$ cujas arestas A representam procedimentos que levam de um estado e_i para um estado e_{i+1}; a resolução do problema torna-se, então, encontrar a sequência de procedimentos que conecte um vértice inicial e_0 no vértice objetivo e_{goal}; o caminho que conecta estes dois vértices será a solução do problema, isto é, a sequência de procedimentos necessários para sair do estado inicial e chegar no estado objetivo; para evitar que caminhos repetidos sejam explorados, o grafo G normalmente é transformado para uma árvore, um tipo de grafo conexo acíclico; é com este tipo de representação que se executa um algoritmo de busca (dos caminhos ou nós solução). Usando este tipo de abordagem diversos problemas podem, a princípio, ser resolvidos pelo computador, como quebra-cabeças, roteamento, navegação de robôs, entre outros [109].

As estratégias de busca podem ser cegas, informadas ou competitivas: estratégias cegas exploram os nós da árvore do problema de maneira sistemática, sem usar nenhum tipo de

[4]Em última instância, todas as representações expressam-se fisicamente por arranjos de bits na memória e são equivalentes entre si, havendo, contudo, vantagens intrínsecas quanto às conveniências e formas de pensar para os desenvolvedores dos algoritmos de IA [103].

[5]Sistemas que manipulam expressões matemáticas analiticamente são denominados de Sistema Algébrico Computacional (do inglês, *Computer Algebra System*, CAS) [113].

[6]Uma das primeiras linguagens de programação amplamente utilizadas na área de IA [104].

informação adicional; estratégias informadas são denominadas de busca heurística e utilizam alguma informação para escolher quais nós explorar primeiro; já buscas competitivas implementam ações adversariais a cada iteração do algoritmo, emulando um jogo. São exemplos de algoritmos de busca cega a Busca em Profundidade (do inglês, *Depth-First Search*, DFS) e a Busca em Largura (do inglês, *Breadth-First Search*, BFS); para a busca heurística temos os algoritmos A* (lê-se A estrela) e Busca Gulosa; como exemplo de busca competitiva podemos citar o algoritmo Minimax.

Assumindo que existe uma solução no espaço de estados, um algoritmo é dito *completo* se for de fato capaz de encontrá-la [107]. Entretanto, dentre os algoritmos completos, nem todos encontrarão a solução ótima. Isto ocorre quando há mais de uma solução, seja mais de um nó objetivo ou mais de um caminho ligando o nó inicial ao nó objetivo. Além disso, nem todos encontrarão a solução no mesmo tempo e usando a mesma quantidade de memória. Há portanto duas dimensões a se considerar além da completude, respectivamente, a otimização da resposta[7] e a complexidade computacional (de memória e de tempo). É importante notar, no pior dos casos, todas essas estratégias são semelhantes a uma busca por força bruta (também chamada de busca exaustiva), isto é, tentar todas as combinações de rotas possíveis. Evidentemente a complexidade algorítmica deve ser uma função do tamanho do espaço de busca, ou seja, do número de nós da árvore. Em alguns casos, a *explosão combinatória* dos estados torna inviável, em tempo e memória, o uso de qualquer uma dessas abordagens, isto é, há mais estados a serem explorados do que memória e/ou tempo computacional disponíveis para a exploração. Por outro lado, a otimização da solução depende do custo de cada procedimento, normalmente representado pelo peso da aresta no grafo. Estratégias heurísticas como a Busca Gulosa priorizam a escolha do próximo estado com menor custo com relação ao atual, enquanto a Busca A* leva em consideração também o custo total do caminho percorrido; para uma árvore com pesos iguais a Busca em Largura é ótima assim como a A* o é para grafos ponderados arbitrariamente. Entretanto ambos A* e BFS são assintoticamente exponenciais em tempo e espaço, o que os torna inviáveis para solução de problemas com espaço de busca muito grande. Neste caso, quando não é necessário armazenar a sequência de procedimentos (ou seja, a solução do problema é um estado no grafo), pode-se aplicar algoritmos de busca local que abrem mão da exploração sistemática, mas consomem menos memória e podem gerar soluções razoáveis mesmo em espaços de busca muito grandes ou contínuos [109]. Método de busca local, também chamados de meta-heurística, funcionam partindo de uma configuração inicial e fazendo consecutivas alterações nessas configurações até convergirem para um estado [107]. Exemplos de algoritmos de busca local são o Subida de Encosta (*Hill-Climbing*), o Recozimento Simulado (*Simulated Annealing*), o Metropolis–Hastings, Algoritmos Genéticos, Otimização por Colônia de Formigas e Redes Neurais Artificiais.

Outro conceito importante na complexidade computacional é a definição da classe do

[7]Nós referimos a solução ótima do problema original, não necessariamente ao algoritmo ótimo, i.e., ao algoritmo que encontra uma solução do problema no menor tempo ou com a menor quantidade de memória.

problema. Considerando recursos de tempo, problemas de decisão[8] (aqueles que podem ser respondidos com *sim* ou *não*) podem ser classificados como P (tempo polinomial) ou NP (tempo não polinomial). Problemas do tipo P podem, em principio, serem resolvidos por Máquinas de Turing Determinísticas em tempo polinomial com relação ao tamanho da entrada; já problemas do tipo NP são verificáveis em tempo polinomial mas sua resolução exige tempo exponencial em uma Máquina de Turing Determinística [115]. Um problema NP só é resolvível em tempo polinomial por uma Máquina de Turing Não-Determinística.

Considerar o tempo de computabilidade de um problema é importante quando consideramos a aproximação pela busca da solução em um espaço de soluções, isto é, quando enumeramos de alguma forma nosso espaço de busca. A explosão combinatória (quantidade de estados ou quantidade de caminhos no grafo do problema) pode tornar inviável para um algoritmo que computa a solução em tempo exponencial com relação a entrada (número de soluções que precisam ser testadas). Mesmo algoritmos que computem em tempo polinomial podem ser marginalmente uteis em espaços de busca realmente grandes. Essa discussão nos leva a outra oportuna definição de AI:

> "Inteligência Artificial é o estudo de técnicas destinadas a resolver problemas exponencialmente difíceis em tempo polinomial, pela utilização de conhecimento a respeito do domínio do problema" – Elaine Rich. [104]

Essa definição, apesar de formalmente imprecisa, traz uma ideia fundamental sobre a real intenção de se utilizar técnicas de IA para resolver problemas dos mais variados tipos. Em última instância estamos propondo um algoritmo de solução do problema que seja melhor em tempo do que a busca exaustiva, sem que necessariamente estejamos convertendo um problema NP em P, já que isso seria equivalente a assumir que existe pelo menos um problema no qual $P = NP$ [9].

A simplicidade da busca exaustiva pode ser ilustrada pela técnica de gere-e-teste: gere uma solução possível, teste-a para verificar se de fato é uma solução. Neste paradigma, é preciso implementar dois módulos, um gerador de soluções e um testador de soluções [103]. Entretanto, isso não significa que testar uma solução em tempo polinomial é garantia de encontrar a solução real em tempo polinomial: o número de soluções possíveis a serem testadas pode ser muito grande. Bons geradores devem ser completos (produzirem todas as soluções possíveis), precisos (sem repetir soluções) e informados (limitam o espaço de busca) [103]. É exatamente nesse sentido que discutimos alguns algoritmos clássicos de IA nos parágrafos antecedentes, quando apresentamos a ideia de busca informada (heurística).

[8] Problemas de otimização e busca também podem ser classificados como P e NP, notando-se que, em geral, é possível obter uma versão de um problema de otimização ou de busca como um problema de decisão. Por exemplo, encontrar a rota que minimiza a distância em um grafo pode ser reduzido a uma versão de decisão na qual pergunta-se: dada uma rota, essa rota tem ou não tem a distância mínima?

[9] Dizer que $P = NP$ ou que $P \neq NP$ ainda é um problema em aberto [116].

3.1.1.2 Explicabilidade e Interpretabilidade

Não obstante, uma discussão diferente que utiliza ideias desenvolvida no âmbito da IA Simbólica é a explicabilidade e interpretabilidade das soluções e algoritmos. Explicabilidade refere-se à capacidade de um modelo de IA fornecer explicações claras e compreensíveis sobre suas decisões, enquanto interpretabilidade está relacionada à compreensão intrínseca do funcionamento interno do modelo, permitindo que humanos entendam a lógica por trás das predições. Embora os termos sejam frequentemente usados de maneira intercambiável, a diferença sutil está na perspectiva. A explicabilidade está mais focada na capacidade de explicar as saídas do modelo, enquanto a interpretabilidade está relacionada à compreensão do funcionamento do modelo em si. Em geral, essa área da IA é conhecida como XAI (do inglês, *Explainable Artificial Intelligence*).

Na IA simbólica, as representações de conhecimento são explicitamente definidas por meio de símbolos e regras, tornando mais fácil compreender como o sistema toma decisões. Contudo, existem limitações em lidar com a complexidade e a incerteza presentes em alguns domínios do mundo real. Sistemas baseados em conhecimento tem um problema intrínseco relacionado à quantidade de conhecimento necessário para tratar prolemas reais: quanto mais conhecimento, mais difícil é acessar a informação relevante, e mais conhecimento deverá ser adicionado para auxiliar na busca, alimentando um ciclo de crescimento infinito [104]. Em contraste, abordagens baseadas em aprendizado de máquina podem oferecer desempenho superior em certos contextos, porém, muitas vezes, sacrificam explicabilidade e interpretabilidade em favor da capacidade de lidar com dados mais complexos e não estruturados. O desafio atual é encontrar abordagens híbridas que combinem as vantagens de ambas as abordagens como as técnicas neuro-simbólicas [112].

3.1.2 Aprendizado de Máquina

Aprendizado de máquina é uma classe de algoritmos de inteligência artificial na qual o problema original é convertido em um *problema de aprendizado* ou se utiliza memória de buscas anteriores como informação inicial. Uma primeira tentativa de definição para essa área foi dada por Arthur Samuel em 1959:

> "Aprendizado de Máquina é o campo de estudo que visa dar aos computadores a capacidade de aprenderem sem ser explicitamente programados" – Arthur Samuel. [117]

A abordagem de Samuel neste artigo [117] é denominada de **aprendizado por hábito**: uma árvore de soluções representa movimentos e posições em um jogo de damas; uma técnica de busca competitiva é usada para escolher o melhor nó a ser explorado a cada iteração; entretanto, como a árvore não pode crescer indefinidamente (espaço de busca muito grande), é necessário parar a busca em determinados momentos; o aprendizado entra guardando em memória uma informação já processada durante uma expansão, de forma que, se o algoritmo de IA se deparar com uma posição inicial igual aquela que gerou

a expansão armazenada, ele não precisará expandir a busca, ou seja, terá aprendido a melhor jogada por já ter jogado aquela posição [104].

Essa é uma abordagem essencialmente simbólica, porém extremamente relevante por destacar a importância de não explicitar a solução do problema durante a criação do algoritmo de AI. Isso ocorre pois nenhum conhecimento sobre estratégias para se jogar damas é codificado em uma base de conhecimento do modelo, exceto as regras de movimento que definem os estados possíveis na árvore. Todo conhecimento das estrategias é *aprendido* pelo algoritmo a cada vez que é executado.

Outra forma de aprendizado é o **aprendizado indutivo** [107]. Nessa abordagem, algoritmos de ML aprendem a induzir uma função ou hipótese capaz de resolver um problema a partir de dados que representam instâncias do problema a ser resolvido [118]. Estes métodos são baseados no princípio de que, se for encontrada uma função capaz de mapear corretamente um grande conjunto de dados de treinamento, então a função também poderá ser usada para mapear dados não observados anteriormente, isto é, será possível generalizar para instâncias não vistas do mesmo problema [107]. Dessa forma, no aprendizado indutivo, espera-se partir de observações de soluções para uma solução geral, diferindo de técnicas de dedução (como em modelos simbólicos baseados em regras): em modelos dedutivos, se as premissas forem verdadeiras, então a solução encontrada é garantidamente correta; em modelos indutivos as conclusões podem estar incorretas [109]. Alguns autores utilizam o termo aprendizado estatístico [30]

Existem diferentes estratégias de aprendizado indutivo:

- **Aprendizado Supervisionado:** o algoritmo é treinado em um conjunto de dados rotulado, onde as entradas (dados de treinamento) estão associadas a saídas desejadas. O objetivo é fazer com que o modelo generalize para novos dados, prevendo corretamente as saídas.

- **Aprendizado Não Supervisionado:** o algoritmo é treinado em um conjunto de dados não rotulado, e o sistema tenta aprender padrões e estruturas nos dados sem orientação explícita sobre as saídas desejadas.

- **Aprendizado Semi-Supervisionado:** o algoritmo é treinado com um conjunto de dados que contém tanto exemplos rotulados quanto não rotulados. Essa abordagem é útil em situações em que rotular grandes quantidades de dados é caro ou demorado. Para melhorar as predições sobre os dados não rótulos, são empregadas técnicas de Aprendizado Ativo (do inglês, *Active Learning*) nas quais um oraculo (ser humano ou outro método computacional) é usado para fornecer o rótulo real das instâncias não rotuladas que produziram os maiores erros de predição. Por outro lado, o aprendizado semi-supervisionado também se refere ao uso de dois algoritmos encadeados, um de aprendizado não supervisionado e outro de aprendizado supervisionado. Neste caso, o algoritmo não supervisionado é utilizado para obter uma nova representação dos dados de entrada e o algoritmo supervisionado para usar essa representação em uma tarefa de classificação ou regressão.

- **Aprendizado por Reforço:** O algoritmo aprende através da interação com um ambiente. Ele recebe realimentação (*feedback*) na forma de recompensas ou penalidades à medida que realiza ações em um ambiente, com o objetivo de maximizar as recompensas ao longo do tempo.

Cada estratégia de aprendizado permite definir uma tarefa que o algoritmo é capaz de realizar. De maneira geral, algoritmos de aprendizado supervisionado podem realizar tarefas preditivas de classificação (problema de decisão binária ou multiclasse, onde as novas instâncias são mapeadas para *strings* que identificam um das classes de treinamento) e de regressão (problema de predizer um valor de saída contínuo para uma nova instância de entrada). Já algoritmos não supervisionados são utilizados para obter novas representações das instâncias de entrada (transformar a instância de entrada em uma nova instância com tipo de dado diferente ou com uma dimensão[10] diferente) ou realizar a tarefa descritiva de agrupamento de dados (atribuir rótulos inteiros as instâncias).

3.1.2.1 Algoritmos de Aprendizado de Máquina

Os algoritmos de ML são construídos através de um módulo de aprendizado com viés indutivo (a representação que descreve a hipótese induzida) [118]. Por exemplo, algoritmos de ML denominados analogistas consideram cada instância de entrada como um ponto em um espaço N-dimensional; em um KNN[11] novas instâncias podem ser classificadas como pertencentes a uma classe pela distância euclidiana entre o novo ponto e todos os pontos usados no treinamento; o algoritmo induz por analogia que se uma nova instância está próxima espacialmente de um conjunto de instâncias que sabidamente pertencem em sua maioria a uma determinada classe, então aquela nova instância também deve pertencer a essa classe. Exemplos de algoritmos analogistas, são o KNN e a SVM[12] (supervisionados) e o K-Means (não supervisionado).

Já Redes Neurais Artificiais (RNA, em inglês ANN de *Artificial Neural Networks*) tem o viés indutivo bioinspirado, baseado no modelo matemático de um neurônio, que pode ou não, estar conectado com muitos outros neurônios. Neste tipo de modelo, o viés indutivo está representado nos coeficientes (peso sinápticos) da RNA que são aprendidos durante o treinamento.

Por outro lado, Árvores de Decisão e algoritmos de ensemble baseado em árvores (como o RandomForest e o GradientBoosting) possuem um módulo aprendiz inspirado na IA simbólica: as instâncias de treinamento são divididas por sequências de regras de decisão do tipo se-senão; a forma como as regras são geradas e encadeadas é determinado por uma métrica de ganho de informação, sendo um processo automático baseado nos próprios

[10] Por dimensão em ML nos referimos a quantidade de atributos (também chamado de *features*) que uma instância do problema tem. Por exemplo, se um átomo é descrito na entrada de um algoritmo de ML apenas pelo seu número atômico e sua eletronegatividade, então instâncias de átomos terão dimensão igual a 2, i.e., serão descritas por dois atributos.

[11] Algoritmo analogista, supervisionado de classificação, denominado de de k-vizinhos mais próximos (do inglês, *k-Nearest Neighbors*, KNN).

[12] Máquina de Vetor Suporte, do inglês, *Support Vector Machine*.

dados de treinamento; como resultado o algoritmo treinado não se difere de uma sequência de se-senão sobre os atributos das instâncias de entrada.

Cada representação utilizada define a preferência ou viés de representação do algoritmo e pode restringir o conjunto de hipóteses que podem ser induzidas pelo algoritmo [118]. A representação assumida pelo módulo de aprendizado também tem impacto na interpretabilidade e explicabilidade do modelo, por exemplo, Árvores de Decisão mais acessíveis em termos de interpretabilidade em comparação com alguns outros algoritmos mais complexos como RNAs.

Em geral os elementos da representação do viés indutivo necessitam de uma função de perda (também denominada função de custo)[13] para traduzir o problema de aprendizado em um problema de otimização ou ranqueamento. Por exemplo, o algoritmo KNN ranqueia os ponto pela distância euclidiana (outras métricas de distância também podem ser usadas); já Árvores de Decisão otimizam o ganho de informação através de uma medida de impureza (como entropia ou coeficiente de Gini) para determinar a sequência de se-senão que melhor classifica os dados; RNAs de camada regressiva podem usar o erro quadrático médio[14] sobre os dados de treinamento para determinar os valores do pesos sinápticos que compõem a função de mapeamento entre os valores da instância de entrada e o valor predito; já RNAs de camadas de classificação normalmente empregam funções de perda diferentes para medir o acerto discreto entre classes, como a entropia cruzada categórica.

Mesmo com uma função custo explicita, não necessariamente o algoritmo de ML terá um modelo matemático explicito de mapeamento entre as entradas e saídas. Por exemplo, apesar do KNN conseguir generalizar uma predição para uma nova instância, ele não faz isso substituindo os valores da nova instância em uma equação aprendida durante o treinamento. Já RNAs aprendem de fato uma equação[15], de forma que a etapa de inferência sobre novas instância é executada calculado-se a equação aprendida com os atributos dessa nova entrada.

Nos algoritmos de aprendizado supervisionado, os momentos de treinamento e inferência são claramente bem definidos. Seja X_{train} uma matriz onde cada n-ésima coluna representa um atributo em um total de N atributos e cada k-ésima linha representa uma instância diferente do problema, a etapa de treinamento é definida como:

[13]Na literatura especializada esses termos podem aparecer de maneira intercambiável ou não. Alguns autores preferem usar o termo função de perda (do inglês, *loss function*) para designar apenas o erro entre o valor predito de uma única instância e seu valor real, e o termo função de custo (*cost function*) para designar a soma dos erros de predição de todas as instâncias, i.e., o somatório da função de perda sobre todos os dados usados.

[14]Isto é, similar a regressão linear, que também pode ser considerada um algoritmo de ML supervisionado. Entretanto, a regressão linear tem solução analítica, enquanto métodos de RNA requerem otimização (minimização do erro) por métodos numéricos.

[15]Mais precisamente eles aprendem os coeficientes da equação; a equação em si está embutida no viés indutivo do modelo, isto é, no modelo matemático de rede neural. Métodos que aprendem equações para mapear os dados são denominados de Regressão Simbólica uma vez que precisam não só aprender os coeficientes da equação, mas também a sequência símbolos de operação (como os algébricos $+, -, \times, \div$).

$$\text{Algoritmo}(\theta) = \underset{\theta}{\text{argmin}} \frac{1}{k} \sum_{i=1}^{k} \mathcal{L}(f(X_{\text{train}}^{(i)}; \theta), y_{\text{train}}^{(i)}), \tag{3.1}$$

onde $\mathcal{L}$ representa a função de perda que mede a discrepância entre as predições do modelo e os rótulos reais y_{train}, $f(X_{\text{train}}^{(i)}; \theta)$ é a predição do modelo com os parâmetros θ para a entrada $X_{\text{train}}^{(i)}$.

E a inferência ocorre como:

$$\hat{y} = f(X_{\text{new}}; \theta) \tag{3.2}$$

onde $\hat{y}$ é valor de saída predito pelo modelo paramétrico $\text{Algoritmo}(\theta) = f(\theta)$.

É importante notar que nem todos os algoritmos de ML são paramétricos. Algoritmos da família das Árvores de Decisão e do KNN, por exemplo, não possuem um parâmetro θ que representa os coeficientes aprendidos, já que o viés indutivo desses modelos não assume uma equação predeterminada. O resultado do treinamento desses algoritmos constitui soluções próprias, por exemplo, Árvores de Decisão aprendem uma sequência de inequações utilizadas para separar as regiões do espaço de entrada, de forma a selecionar um nova entrada para uma região específica que representa uma classe de maioria no treinamento ou que recebe o valor médio dos exemplos de treinamento; o KNN resulta de maneira semelhante em classe de maioria e valor médio, mas não aprende uma regra, tendo tais atribuições sendo realizadas durante sua execução[16] por cálculo de distância espacial do novo dado de entrada e dos dados de treinamento.

Além dos parâmetros aprendíveis (acima discutido com a notação θ), a maioria dos algoritmos de ML possuem hiperparâmetros específicos relacionados aos próprios métodos e cujos valores não são modificados durante o treinamento. Estes hiperparâmetros, portanto, não são aprendíveis pelo viés indutivo, e devem ser selecionados antes do treinamento do modelo, servindo para realizar ajustes finos no comportamento do algoritmo. Por exemplo, o KNN possui um hiperparâmetro k que define quantos pontos vizinhos serão considerados durante a decisão de maioria para a classificação; Árvores de Decisão têm um hiperparâmetro da profundidade máxima que uma árvore pode ser expandida durante o treinamento; RNAs possuem hiperparâmetro do número de épocas de treinamento devido seu treinamento ser realizado por retropropagação quando são usadas muitas camadas de neurônios (o próprio número de neurônios em cada camada da rede é em si um hiperparâmetro para esses algoritmos).

O ajuste de hiperparâmetros de um algoritmo de ML também é denominado de etapa de validação, onde um conjunto de dados de treinamento e de teste é usado para escolher quais são os melhores hiperparâmetros através de uma amostragem sistemática (conhecida como GridSearch) ou aleatória (conhecida como RandomSearch). Aqui é extremamente importante separar os dados usados na validação daqueles que de fato serão usados para o treinamento do modelo, seja impedindo que se repitam por exclusão ou utilizando estraté-

[16] Um algoritmo de aprendizado preguiçoso (em inglês, *lazy learning*) é aquele que adia a generalização dos dados de treinamento até o momento da predição, realizando pouco ou nenhum processamento durante o treinamento e armazenando os dados para uso posterior durante a inferência.

gias de amostragem como a validação cruzada [119]. Isso ocorre para garantir generalidade ao modelo aprendido, uma vez que a escolha dos melhores hiperparâmetros usando o mesmo conjunto de dados faria o modelo ficar enviesado para aquele conjunto de dados (o que é denominado sobreajuste ou *overfitting*).

Uma forma de saber se o modelo está sofrendo de *overfitting* é separar aleatoriamente o conjunto de dados inicial (antes do ajuste de hiperparâmetros) em, pelo menos, dois conjuntos: dados de treinamento e dados de teste. Neste esquema, os dados de treino são usados por validação cruzada para selecionar os melhores hiperparâmetros do algoritmo, e em seguida o algoritmo é treinado com o conjunto total de dados de treinamento; na etapa de inferência são usados os dados de teste dos quais já é conhecida a saída; compara-se uma métrica de desempenho (por exemplo, o erro médio absoluto ou a acurácia) para o treinamento e para o teste; se o desempenho de teste for muito pior do que o desempenho de treinamento, então o modelo estará apresentando *overfitting*. Aqui, assume-se que os dados de treinamento, de teste, e de validação partem da mesma distribuição original (por isso a amostragem aleatória durante a separação). Problemas de generalização também podem ocorrer quando as novas instâncias a serem inferidas estão muito fora da distribuição de treinamento (do inglês, *out-of-distribution*, OOD) [112].

3.1.2.2 Representação de Dados

A representação da entrada também é uma característica a ser projetada nos algoritmos de ML. Em geral, a representação mais simples assumida pela maioria dos algoritmos é a representação vetorial, onde cada entrada é um vetor de números reais. Outras representações, em geral, exigem modificações na arquitetura interna do algoritmo. Imagens, áudios, objetos de texto (*strings*) e grafos são exemplos de outros formatos de dados que podem ser entrada de algoritmos de ML, mas que necessitam de arquiteturas específicas. Contudo, devemos destacar que cada uma dessas representações pode ser transformada em uma representação vetorial (ou, de maneira mais geral, matricial) através de algum pré-processamento específico. A distinção reside na localidade dessa etapa em si. Por exemplo, em uma rede neural convolucional treinada para a tarefa de classificação de imagens, as entradas são tensores $NxMx3$ onde N e M são os números de *pixels* e 3 a quantidade de canais de cor (três para o padrão RGB); as camadas convolucionais da rede são construídas de maneira a transformar os dados iniciais em vetores planificados antes de serem conduzidos a uma camada totalmente conectada similar a uma *multilayer* perceptron; em outras palavras, podemos enxergar as camadas convolucionais dessa rede como etapa de pré-processamento que transformam os dados para representação vetorial - ou seja, a transformação é interna a arquitetura do algoritmo de IA em si. Por outro lado, poderíamos criar um algoritmo de pré-processamento independente, que, por exemplo, calculasse características específicas da imagem, como frequência de *pixels* de uma determinada intensidade para um determinado canal, a fim de formar um vetor de *features* que representem a imagem. Neste caso também estaríamos transformando a representação da imagem como

no casos anterior, mas a transformação estaria sendo feita fora do algoritmo de IA. No primeiro caso dizemos que a Extração de *Features* foi realizada pelo próprio algoritmo de IA, enquanto nos segundo dizemos que foi realizada uma Engenharia de *Features* para representar o problema baseada em conhecimento de domínio do especialista.

Os algoritmos de ML também são categorizados de acordo com essa habilidade de aprender a extrair *features*. É comum a denominação Aprendizado Raso (do inglês, *Shallow Learning*) para modelos que necessitam de engenharia de*features*e de Aprendizado Profundo (do inglês, *Deep Learning*) para modelos que implementam a extração de *features* na própria arquitetura [120]. Diz-se desses últimos que eles são inerentemente capazes de aprender a representação importante dos dados já que não são deliberadamente selecionadas características dos dados brutos. Deve-se notar, contudo, que boa parte da "inteligência"presente nesse aprendizado reside na própria arquitetura, uma vez que essa arquitetura considera em si a natureza específica do dados brutos que serão processados (i.e., se serão imagens, áudios, texto, etc). Não obstante, lembramos que essa categorização não é unicamente aceita, com outros autores incluindo na categoria de Aprendizado Profundo, algoritmos de ensemble como RandomForest e GradientBoosting, em especial quando esses algorítimos utilizam várias níveis de ramificação das árvores exploradas (de onde se entende a "profundidade").

3.1.2.3 Dados Sintéticos e Aumento de Dados

Por dados sintéticos nos referimos a dados que não foram gerados por eventos da realidade, como dados capturados por sensores e outros instrumentos de medição ou gerados pela atividade humana[17]. Logo, um dado sintético é uma informação gerada artificialmente, em vez de ser extraída diretamente de observações reais. Neste sentido, resultados de simulações podem ser entendidos como dados sintéticos.

Já o aumento de dados (do inglês, *data augmentation*[18] é a técnica de gerar variações artificiais dos dados de treino para aumentar a diversidade e robustez de um modelo treinado com tais dados.

Existem diversas formas de criar dados sintéticos por *data augmentation*. Para imagens, podemos criar versões rotacionadas, transladadas, escalonadas, com cortes, adição de ruído, alteração de brilho, contraste ou cor [121]. Em áudio, usam-se variações de pitch, tempo, adição de ruído, corte ou mixagem com outros sons [122]. Já para dados tabulares, as operações de aumento precisam ser feitas no próprio espaço de *features*. Para todos esses exemplos são realizadas alterações no dado em sua forma original, criando um novo dado que está ligeiramente diferente do original.

Por outro lado, podemos realizar aumento da dados através de sobreamostragem (do inglês, *oversampling*) dos exemplos de treinamento. A sobreamostragem com reposição (do inglês, *oversampling with replacement*) envolve duplicar o dado cuja característica pre-

[17]Isto é, dados de origem social, como demografia, economia, etc.

[18]Por rigor, optamos pela tradução literal "aumento de dados". Entretanto, na comunidade de inteligência artificial lusófona é frequente o uso do neologismo "*augmentação* de dados".

tendemos aumentar em importância para o modelo (normalmente as classes minoritárias). Esses dados são selecionados de maneira aleatória para serem duplicados desde que possuam a característica a ser aumentada.

Também podemos fazer a sobreamostragem usando algum método de alteração do dado original. A técnica de SMOTE (*Synthetic Minority Over-sampling Technique*), por exemplo, faz uma sobreamostragem gerando dados sintéticos no espaço de *features* de vez de operar puramente do espaço de dados [123]. Especificamente, o SMOTE seleciona exemplos aleatórios da classe minoritária e, para cada um, escolhe um de seus k-vizinhos mais próximos também pertencente à classe minoritária. Em seguida, ele cria uma nova amostra sintética interpolando linearmente entre o ponto original e o vizinho escolhido, gerando assim exemplos que são convexas combinações no espaço vetorial das *features*. Essa abordagem mitiga problemas de sobreajuste (*overfitting*) frequentemente associados à replicação direta de exemplos minoritários e ajuda a expandir a região de decisão da classe minoritária, promovendo uma melhor generalização do modelo.

3.1.2.4 Avaliação de Desempenho

Algoritmos de aprendizado de máquina precisam ser testados quanto a seu desempenho em relação a tarefa aprendida. Para os algoritmos supervisionados os testes podem ser conduzidos de acordo com o gabarito observado, isto é, os próprios dados utilizados. Para assegurar que o modelo tenha capacidade de generalização, a avaliação de desempenho ocorre com a divisão do conjunto de dados entre treinamento e teste, de forma que os testes propriamente ditos são conduzidos apenas com os dados desse respectivo conjunto.

Além dessa questão metodológica, é necessário definir uma métrica de avaliação. Cada tarefa terá sua métrica de acordo com a natureza do dado (discreto ou contínuo, por exemplo). Métricas também tem suas próprias característica, de maneira que é comum utilizar mais de uma métrica para medir o desempenho de um mesmo sistema, a fim de eliminar vieses contidos nas definições das próprias métricas.

Regressão

Para a tarefa de regressão é comum a utilização do Erro Absoluto Médio (do inglês, *Mean Absolute Error*, MAE), definido como:

$$MAE = \frac{1}{k}\sum_{j=1}^{k}\left|y_{true_j} - y_{pred_j}\right|, \tag{3.3}$$

onde k é a quantidade de exemplos usados como teste, y_{true} o valor real esperado e y_{pred} o valor predito pelo método.

Para expressar a capacidade preditiva de um método de ML é comum utilizar a razão entre a MAE e o desvio absoluto médio (do inglês, *Mean Absolute Deviation*, MAD). A MAD é definida como:

$$\text{MAD} = \frac{1}{k}\sum_{i=1}^{k}|y_{true_j} - \bar{y}_{true}|, \tag{3.4}$$

onde $\bar{y}_{true}$ é a média sobre todos os valores esperados.

A ideia subjacente a essa comparação é que, se fossemos utilizar um sistema preditivo rudimentar, poderíamos supor que o valor alvo é o valor da média dos valores de treinamento. Logo, a diferença da média que pode ser medida pelo desvio padrão ou pelo desvio absoluto médio medem uma espécie de erro esperado para esse modelo preditivo. Ao comparar a MAE com a MAD estamos então comparando quão superior o método preditivo foi com relação ao método de "predizer pela média". De maneira geral, para razões $MAD : MAE > 5$ diz-se que o modelo tem boa capacidade preditiva [124].

A MAE é uma métrica fácil de ser avaliada uma vez que o resultado obtidos está na mesma unidade do valor predito. Para mensurar valores absolutos que podem depender especificamente da tarefa realiza, a razão entre MAE e MAD carateriza uma forma normalizada, constituindo um número adimensional.

Um problema que surge do uso dessa métrica tem relação com o fato da operação módulo realizada internamente nas Eq. 3.3 e 3.4, possuir uma descontinuidade. Para algumas considerações adicionais, seria interessante calcular a derivada do erro, de forma que essa descontinuidade pode trazer complicações computacionais. Uma proposta auxiliar é o cálculo das médias dos quadrados dos erros, conhecida em inglês como *Mean Squared Error* (MSE). Neste caso, temos que:

$$MSE = \frac{1}{k}\sum_{j=1}^{k}\left(y_{true_j} - y_{pred_j}\right)^2. \tag{3.5}$$

Nota-se que o resultado da MSE corresponderá a um valor normalmente maior do que os valores observados com a MAE, uma vez que tais valores estarão ao quadrado. Assim, a unidade da MSE corresponde ao quadrado da unidade da variável sendo predita. Para retornar a unidade original é comum empregar-se a raiz quadrada de MSE, conhecida em inglês como *Root Mean Squared Error* (RMSE):

$$RMSE = \sqrt{\frac{1}{k}\sum_{j=1}^{k}\left(y_{true_j} - y_{pred_j}\right)^2}. \tag{3.6}$$

Novamente é possível realizar uma normalização pelo método preditivo da média. Dessa vez utiliza-se o desvio padrão (em inglês, *Standard Deviation*, STD, também normalmente denotado por σ):

$$STD = \sqrt{\frac{1}{k}\sum_{j=1}^{k}\left(y_{true_j} - \bar{y}_{true}\right)^2}. \tag{3.7}$$

Da mesma forma que para os valores absolutos calculamos MAD/MAE, para as raízes

dos quadrados calculamos STD/RMSE, quando queremos utilizar uma métrica normalizada adimensional.

Outra métrica oriunda da Estatística é o Coeficiente de Determinação, denotado por R^2:

$$R^2 = 1 - \frac{\sum_{j=1}^{k} \left(y_{true_j} - y_{pred_j}\right)^2}{\sum_{j=1}^{k} \left(y_{true_j} - \bar{y}_{true}\right)^2}. \tag{3.8}$$

Essa métrica também é adimensional, tendo seu valor máximo em 1: se o erro dos resíduos for zero (o numerador na Eq.3.8), então o modelo terá predito exatamente os valores observados e $R^2 = 1$. Diferente das métricas absolutas MAE, MSE e RMSE que buscamos o menor valor (quanto mais próximo de 0, melhor), para o R^2 valores próximos de 0 são extremamente ruins. São considerados bons modelos $R^2 \geq 0.95$, porém em casos específicos modelos com $R^2 \geq 0.8$ já são considerados aceitáveis.

Destaca-se também a diferença entre a métrica R^2 e a métrica STD/RMSE. Enquanto a primeira está medindo uma relação inversa entre os quadrados dos erros e variância dos dados, a segunda mede quantas vezes o erro é menor do que o desvio padrão do dados.

Uma possível extensão de R^2 também é utilizada para diminuir um efeito indesejado da quantidade de variáveis do modelo: conforme mais *features* são usadas, maior será o R^2, o que não deveria necessariamente indicar um modelo melhor. O Coeficiente de Determinação Ajustado $\bar{R}^2$ pode ser definido como [125]:

$$\bar{R}^2 = 1 - \frac{\sum_{j=1}^{k} \left(y_{true_j} - y_{pred_j}\right)^2}{\sum_{j=1}^{k} \left(y_{true_j} - \bar{y}_{true}\right)^2} \frac{(k-1)}{(k-n-1)}, \tag{3.9}$$

onde k é é o número de amostras e n é o número de variáveis do modelo.

Classificação

Como métrica de desempenho da tarefa de classificação são definidas a acurácia, a precisão e a revocação:

$$Acurácia = \frac{|y_{pred} \cap y_{true}|}{|y_{pred}|} = \frac{VP + VN}{VP + VN + FP + FN}, \tag{3.10}$$

$$\text{Precisão} = \frac{VP}{VP + FP}, \tag{3.11}$$

$$Revocação = \frac{VP}{VP + FN}, \tag{3.12}$$

onde os verdadeiros positivos (VP) contabilizam concordâncias entre os dois métodos no escopo da classificação, os verdadeiros negativos contabilizam concordâncias entre os dois métodos para as instâncias fora do escopo da classificação; e os falsos positivos (FP) e os falsos negativos (FN) contabilizam discordâncias entre os métodos, sendo o primeiro para classificações erroneamente dentro do escopo, e o segundo para classificações erroneamente

fora do escopo.

A acurácia mede a taxa de acerto total. Essa métrica, apesar de intuitiva, pode gerar equívocos em cenários de dados muito desbalanceados (muitos exemplos para uma classe e poucos para outra). Por exemplo, considere um cenário no qual existem apenas 100 amostras positivas em um conjunto de teste de 10000 amostras: se fizermos um modelo ingênuo que classifica pela maioria, teremos acertado 9900 em 10000 testes, obtendo uma acurácia de 99%. Precisão e revocação ajudam a medir, respectivamente, quão preciso foi o modelo em classificar somente exemplos dentro do escopo quando realmente pertencem ao escopo e de não deixar de classificar exemplos dentro do escopo. No exemplo discutido neste parágrafo, tanto a precisão quanto a revocação teriam dado 0% evidenciando a inutilidade do classificador de maioria proposto.

A diferença entre a precisão e a revocação está na natureza do erro que cada uma contabiliza: a precisão está em um erro de seleção, predizer para a classe no escopo uma instância que não pertence a classe; já a revocação o erro está em não predizer para a classe no escopo uma instância que pertence a essa classe. A escolha entre qual tipo de erro é melhor ou pior depende do problema que está sendo resolvido. Porém, em situações em que haja uma relação de compromisso entre essas duas métricas (maximizar um acaba minimizando o outro), então é aconselhável utilizar a métrica F_1-*score*. Essa métrica é definida pela média harmônica entre a precisão a revocação:

$$F_1\text{-score} = \frac{2}{\frac{1}{\text{Precisão}} + \frac{1}{\text{Revocação}}} = \frac{2VP}{2VP + FN + FP}. \tag{3.13}$$

De maneira mais geral, também pode ser utilizado o F_β-*score*, onde β determina a importância dada a revocação:

$$F_\beta\text{-score} = \frac{\beta^2 + 1}{\frac{1}{\text{Precisão}} + \frac{\beta^2}{\text{Revocação}}} = \frac{(1+\beta^2)VP}{(1+\beta^2)VP + \beta^2 FN + FP}. \tag{3.14}$$

Um complicador associado ao uso da precisão, revocação, F_1-*score* e F_β-*score* é o fato dessas métricas serem calculadas com relação ao escopo do modelo, isto é, a uma classe alvo específica. Ou seja, haverá um valor para cada uma dessas métricas para cada classe que o classificador consiga classificar. Logo, o critério exige uma decisão consciente do experimentador sobre qual classe é mais importante no problema que está sendo resolvido. Uma alternativa a priorização de uma classe é o cálculo da média sobre todas as classes. Por exemplo, se tivermos um problema com duas classes, digamos A e B, de forma que $\text{Precisão}_A = 0.8$ e $\text{Precisão}_B = 0.7$, podemos assumir a precisão do modelo como a média aritmética $0.5(0.7 + 0.8) = 0.75$. Por outro lado, as classes A e B podem estar desbalanceadas no conjunto de teste, isto é, termos mais exemplos de teste de uma classe do que outra. Neste caso poderíamos optar por calcular a média aritmética das classes ponderada pelos exemplos de cada uma. Em todos os casos é importante entender como o desbalanceamento das classes pode interferir nas métricas associadas para que as interpretações dos resultados sejam corretas [126].

Outra métrica de desempenho de classificação é o coeficiente κ de Cohen que mede quanto o classificador em questão difere de um classificador que apenas sorteia a classe predita. Esse coeficiente é definido no intervalo $-1 \leq \kappa \leq 1$, sendo valores abaixo de 0.2 considerados próximos ao classificador aleatório enquanto valores próximos de 1 correspondem a um classificador que está predizendo exatamente os valores esperados [127].

3.1.2.5 Meta-Aprendizado

De um ponto de vista metodológico, usar métodos de Aprendizado de Máquina envolve quatro elementos fundamentais: quantidade de dados disponíveis para o modelo induzir o aprendizado; a representação dos dados para o modelo, isto é, o espaço de *features*; os algoritmos e seus modelos assumidos *a priori*; os hiperparâmetros dos algoritmos.

Os dois primeiros itens especificam a quantidade e qualidade de informação disponível para um modelo, enquanto os dois últimos especificam a forma pela qual o modelo (o algoritmo) irá operar sobre esses dados.

Se buscamos o melhor sistema para uma determinada tarefa, será necessário otimizar ao longo dessas quatro dimensões — o que, evidentemente, não é uma tarefa trivial. Uma abordagem possível é utilizar um segundo nível de aprendizado, por meio de algoritmos de ML que predizem qual modelo ou configuração tende a ter melhor desempenho. Isso constitui um dos objetivos centrais do meta-aprendizado.

A princípio, não existe um único algoritmo de ML com o melhor desempenho para todos os casos, já que cada viés indutivo se adapta melhor a determinados tipos de *dataset* (informação disponível); entretanto, durante o processo de aprendizado, é possível aprender qual algoritmo é mais apropriado para aquele *dataset* [118].

De maneira prática, a meta-aprendizagem pode envolver a construção de *features* sobre os *dataset* e sobre os algoritmos. Por exemplo, podemos dizer que um determinado *dataset* é composto 20 *features*, contendo 1000 exemplos distribuindo igualmente entre 5 classes. Outro *dataset* tem 500 exemplos em 3 classes com 5 *features*. O conjunto de vários *datasets* especificados dessa forma compõem em si um novo *dataset* representado por 3 *features* (número de exemplo, número de classes, número de *features*). Se para cada exemplo de *dataset* no nosso *meta-dataset* adicionarmos uma coluna com o nome do algoritmo que obteve a melhor performance de classificação usando aquele *dataset*, então poderemos treinar um meta-algoritmo de ML que tentar classificar para uma tripla ordenada (número de exemplo, número de classes, número de *features*) qual é o melhor classificador a ser usado. Podemos expandir essa análise para incluir características do algoritmo usado em cada *dataset* isolado, como os hiperparâmetros usados.

Observamos que a seleção de *features* não é em si uma etapa de meta-aprendizagem, porém pode ser considerada intrínseca a representação dos dados que são usados por algoritmos de meta-aprendizagem. Se usamos um método *wrapper* de seleção de *features*, então a quantidade e qualidade das *features* estão otimizadas para quele algoritmo de *wrapper* utilizado. Logo, existem espaços de *features* específicos para o melhor desempenho do al-

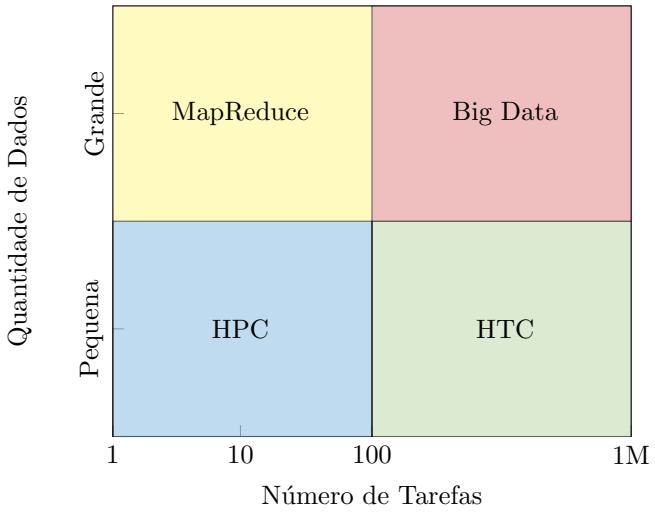
goritmo que podem ou não estar devidamente representados no *meta-dataset*. O mesmo ocorre para a otimização de hiperparâmetros: enquanto a seleção de *features* é uma otimização da informação que um modelo irá receber, a seleção de hiperparâmetros é uma otimização dos modelos que cada algoritmo pode induzir. Destacamos que em ambos os casos, uma rigorosa definição metodológica é importante para assegurar a reprodutibilidade e usabilidade dos resultados obtidos.

3.2 Informática de Materiais

3.2.1 Simulações High Throughput

High Troughput (HT) é um conceito amplo que perpassa por vários domínios. No domínio da Biologia, mais precisamente da farmacologia e design de remédios, o *High Troughput Screening* (HTS) é o uso de equipamentos automatizados para testar rapidamente de milhares a milhões de amostras [128]. Já a *High Troughput Computation* (HTC) caracteriza-se por realizar um grande número de tarefas em uma quantidade pequena de dados. A ideia central em ambos os casos é a alta vazão. Na Figura 3.2 apresentamos um esquema de categorização dos paradigmas tecnológicos para processamento de dados que contempla os métodos HTC:

Figura 3.2: Terminologias das tecnologias utilizadas para processar diferentes volumes de dados.



Fonte: adaptado de [129].

É comum, em se tratando de simulações computacionais de materiais, a necessidade de utilizar infraestrutura de computação de alto desempenho (*High Performance Computing*, HPC). Para cálculos DFT, pode haver um grande esforço computacional (de memória e processamento) que exige tais recursos. Entretanto, a quantidade de dados tratados como um todo é pequena, não excedendo, normalmente, a capacidade de armazenamento de um

computador comum; Também podemos tratar cada simulação DFT como uma única tarefa. Por outro lado, a ideia de HT-DFT é realizar um grande número de simulações (de centenas a milhares) da maneira mais automática possível. Em ambos os casos, a quantidade de dados envolvida é moderada não excedendo alguns gigabytes. A principal diferença é que, enquanto em HPC o paralelismo *intracluster* das subtarefas (por exemplo, dentro da própria simulação) é fortemente acoplado, em HTC o paralelismo *intracluster* ou *intercluster* das tarefas (por exemplo, entre diferentes simulações) é fracamente acoplado. Em outra palavras, o paralelismo de um cálculo DFT realizado em um ambiente HPC representa o uso eficiente dos recursos (tempo de processamento) para uma única simulação, enquanto o paralelismo em HTC refere-se ao uso eficiente dos recursos (tempo de processamento e memória) ao realizar várias simulações independentes concomitantemente.

Quando o volume dos dados aumenta, é necessário empregar técnicas de computação distribuída como o MapReduce, que paraleliza o processamento entre vários *clusters* de computação, de forma que os dados são processados em lotes [100]. Quando a quantidade de tarefas aumenta, assim como a quantidade de dados, estamos em um cenário de *Big Data*, onde combinações das soluções anteriores, acrescidas de outras soluções (como as necessárias para processamento em tempo real) são utilizadas. Estes dois últimos paradigmas não são comuns, ainda, à Informática de Materiais.

O cálculo HT para simulações de materiais data de 1995 para descoberta combinatória de supercondutores [130]. Além das características computacionais já discutidas, métodos HT diferenciam-se de uma abordagem de muitos cálculos pela filosofia de extrair correlações de propriedades que possam ser usados para orientar a busca por sistemas com propriedades específicas para uma finalidade [131]. Por exemplo, pode-se estudar ligas Heusler ferromagnéticas calculando as propriedade de centenas a milhares de estruturas iniciais obtidas ou selecionadas através de alguma heurística, e cuja estrutura eletrônica pode ser determinada através da metodologia DFT [77]. Há duas abordagens principais: criar um conjunto de dados amplo que será filtrado após o término de todos os cálculos (abordagem denominada de *screening* e que será discutida na seção 3.2.4.1); ou escolher a próxima simulação baseada na informação das simulações previamente realizadas ou em uma heurística de seleção (estas abordagens podem ser valer de técnicas de inteligência artificial, tema que será abordado na seção 3.1). Para esta segunda abordagem podemos citar o COPEX, um algoritmo co-evolucionário que realiza HT-DFT em lotes, selecionado melhores estruturas a cada geração [132], como exemplo de HT combinado com técnicas de AI; e o trabalho de *Wang et al.* de substituições de elementos químicos usando como heurística a escala química de Pettifor modificada [133, 134].

Tradicionalmente, para se realizar uma simulação DFT é necessário um software que implementa os cálculos quânticos, dados da estrutura do material a ser estudado e dados sobre os potenciais atômicos; parâmetros de entrada devem ser estabelecidos para determinar o tipo de simulação, por exemplo, se haverá relaxação iônica ou só dos elétrons e como será amostrado o espaço reciproco; estes dados de entrada são inseridos manualmente, e o cálculo é inicializado, normalmente, em um *cluster* HPC. Na metodologia HT-DFT, um

grupo de estruturas é selecionado, os parâmetros de entrada são gerados para os cálculos de todas as estruturas e uma rotina programada gerencia a submissão de cada cálculo para o *cluster* HPC, seja paralelamente, serialmente, ou uma mistura de ambos. Com isso, há uma alta vazão dos resultados (estruturas simuladas), que podem ser explorados por técnicas de mineração de dados, inteligência artificial e otimização para estabelecer quais novas estruturas devem ser simuladas ou quais das estruturas simuladas originalmente possuem as melhores propriedades de interesse.

Para garantir alta vazão é comum aplicar técnicas de baixa fidelidade (*lower-fidelity calculations*), como métodos PBE-GGA. A medida que mais correções e informações são usadas nos métodos DFT, como meta-GGA e funcionais híbrido, os cálculos aumentam a precisão (*higher-fidelity calculations*) ao custo de consumo computacional. Por exemplo o custo computacional do SCAN e r2SCAN é cerca de quatro a cinco vezes o custo do PBE tradicional [135].

Além do arcabouço conceitual, é preciso destacar as principais ferramentas utilizadas no contexto HT-DFT. O VESTA é uma ferramenta utilizada principalmente para visualização de estruturas cristalinas e moléculas [136]. O pymatgen (*Python Materials Genomics*) é uma biblioteca Python[19] que provê representações e funções de alto nível para tratar cristais e moléculas [137]. Assim como o pymatgen, o ASE (*Atomic Simulation Environment*) é uma biblioteca em Python com interface para os principais softwares de simulação de materiais como VASP, Quantum ESPRESSO, SIESTA e LAMMPS [138]. Já o VASPKIT é uma ferramenta auxiliar na análise de dados focado no VASP [139]. Com uma proposta diferente, o AiiDA (*Automated Interactive Infrastructure and Database for Computational Science*) é um *framework* mais robusto que permite configurar fluxos de tarefas em *clusters* HPC, garantindo a procedência de um grande volume de dados e o rastreamento de um grande número de tarefas, de maneira centralizada [140]. Todas essas ferramentas possuem uma comunidade de desenvolvimento ativa, o que garante a implementação de novos recursos e uma manutenção continuada, elementos fundamentais para um bom ecossistema de desenvolvimento de software.

Entre as tecnologias amplamente utilizadas em HPC está a MPI (*Message Passing Interface*), uma biblioteca padrão para paralelização de tarefas em sistemas de memória distribuída, essencial para a execução de simulações em larga escala [141]. Complementarmente, o SLURM (*Simple Linux Utility for Resource Management*) atua como um sistema de gerenciamento de filas e alocação de recursos em clusters, permitindo o agendamento eficiente de tarefas computacionais [142].

No contexto de HTC, destaca-se o HTCondor, um sistema projetado para gerenciamento de grandes volumes de tarefas independentes, comum em experimentos com *workflows* automatizados e reprodutíveis [143]. Já para aplicações que seguem o paradigma MapReduce, o Apache Hadoop oferece uma estrutura robusta, combinando programação distribuída com o HDFS (*Hadoop Distributed File System*) [144]. No tratamento de grandes volumes de dados com maior velocidade e interatividade, o Apache Spark surge como

[19] Python é uma linguagem de programação alto nível, de propósito geral, interpretada e multiparadigma.

uma alternativa moderna ao Hadoop, oferecendo processamento distribuído em memória e suporte a tarefas complexas como aprendizado de máquina e análise de grafos [145].

3.2.2 Bases de Dados

Assim que os resultados das simulações HT-DFT estão disponíveis, eles precisam ser armazenados. O armazenamento pode ser dos dados brutos ou apenas das propriedades de interesse. Neste último caso, esse grande volume de dados deve ser pós-processado para que as propriedades sejam extraídas. A Informática de Materiais então contribui com a sistematização e as ferramentas para construção de bases de dados que armazenem esses resultados.

Exemplos de bases de dados de materiais computacionais são o *Materials Project* [146] e o OQMD (*Open Quantum Materials Database*) [147]. Exemplo de base de dados de materiais experimentais é o ICSD (*Inorganic Crystal Structure Database*) [148]. Na Tabela 3.1 reproduzimos algumas das principais bases de dados disponíveis na área de materiais e uma breve descrição do seu propósito majoritário (algumas bases são multipropósito, contendo dados mais heterogêneos; destacamos a fonte majoritária dos dados).

Tabela 3.1: Bases de dados na área de materiais.

Base de Dados	Descrição
ICSD	materiais inorgânicos, experimental
COD	inorgânicos e orgânicos, experimental
Citrination	experimental e computacional
AFLOW	computacional
Materials Project	computacional
OQMD	computacional
NOMAD	múltiplas fontes
Materials Cloud	múltiplas fontes
OMDB	materiais orgânicos
C2DB	materiais 2D
Materials Web	materiais 2D

Fonte: adaptado de [19]

Um conceito importante na construção dessas bases de dados é conhecido pelo acrônimo FAIR (do inglês, *findable, accessible, interoperable and reusable*). FAIR é uma diretriz que dita que os dados precisam ser encontráveis, acessíveis, interoperáveis e reutilizáveis [19]. Porém há outros elementos fundamentais que devem ser considerados na engenharia das bases de dados e de datasets[20]: a coleta de dados (*Data Collection*), curadoria do *dataset*

[20]Por *dataset* nos referimos a um conjunto de dados, estruturados ou não, que compartilham entre si certa afinidade, normalmente associada à tarefa a qual se aplicam.

(*Dataset Curation*) e o conhecimento do *dataset* (*Dataset Knowledge*) [149]. É preciso prover suporte a essas capacidades para garantir a utilidade dos dados.

No caso da HT-DFT a coleta de dados é resultado da geração intensiva de dados por máquinas, devidamente programadas para realizar os cálculos físicos da simulação dos materiais de interesse. Outras formas de coleta de dados para construção de *datasets* do domínio de materiais serão discutidas na seção 3.2.4. Por outro lado, a questão da curadoria de dados é mais ampla, envolvendo a definição da ontologia e do modelo de dados necessário para representar a informação pretendida. É nesta modelagem que é definido se a base de dados irá armazenar dados brutos ou pós-processados, quais propriedades dos materiais serão armazenadas, como as entrada da base se relacionam entre si, entre outras características relacionadas diretamente ao domínio da base de dados (ciência de materiais), mas também as restrições técnicas (sistemas de informática utilizados). A curadoria também preza pela manutenção da base de dados ao longo do tempo, mantenho sua qualidade e as especificações FAIR.

Do ponto de vista tecnológico, para garantir acessibilidade, interoperabilidade e reutilizabilidade, as bases de dados precisam prover interfaces de programação de aplicações (API, do inglês, *application programming interface*) que permitam integração com rotinas programadas. Em última instância, tais recursos são elementos da web semântica, como aqueles que permitam que os dados sejam lidos por máquinas [150]. Este tipo de recurso permite maior grau de automatização das tarefas. No contexto de materiais, existem esforços para padronização das APIs de acesso as bases de dados, como a iniciativa OPTIMADE (*Open Databases Integration for Materials Design*) [151]. O OPTIMADE estabelece especificações para as APIs, que devem seguir o protocolo REST[21] com acesso aos recursos por URL[22] e respostas via JSON[23].

Deve ficar claro, contudo, que bases de dados e *datasets* não são, necessariamente, o mesmo objeto. Em geral, *datasets* possuem dados que compartilham alguma propriedade entre si ou que especificam uma tarefa de interesse. Existe, portanto, certa homogeneidade nos dados de um *dataset*. Já uma base de dados pode ou não ser homogênea, tendo dados mais diversos, seja nas propriedades ou nos formatos. Por exemplo, um conjunto de dados de materiais bidimensionais ferromagnéticos e antiferromagnéticos pode compor um *dataset* de interesse, sendo extraídos de uma base de dados mais ampla.

3.2.3 Ontologias e Metadados

Metadados são dados sobre dados, isto é, uma informação que define do que se trata um conjunto de dados. Por exemplo, em um *dataset* de materiais no formato de uma tabela de

[21]REST (Transferência de Estado Representacional, do inglês, *Representational State Transfer*) é um estilo de arquitetura para desenvolvimento de serviços web que utiliza os princípios e protocolos da web, especialmente o HTTP, para permitir comunicação entre sistemas de forma simples, escalável e sem estado.

[22]URL (Localizador Uniforme de Recurso, do inglês, *Uniform Resource Locator*), é o endereço de rede onde se encontra algum recurso computacional.

[23]JSON (acrônimo de *JavaScript Object Notation*), é um formato, de padrão aberto independente, de troca de dados simples e rápida entre sistemas, que utiliza texto legível a humanos, no formato atributo-valor (natureza auto-descritiva).

propriedades, os nomes das colunas dessa tabela definem as características desses materiais que são armazenadas. Os nomes dessas colunas são, em si, metadados, já que especificam do que se tratam os dados naquela coluna específica. Porém, metadados não são só essas informações imediatas contidas nas especificações, sendo também informações sobre, por exemplo, como os dados foram gerados, quando foram gerados e quem os produziu.

Ontologia, por sua vez, é uma representação formal e explícita de um conjunto de conceitos dentro de um domínio específico e dos relacionamentos entre esses conceitos. Ela atua como um modelo que descreve a estrutura do conhecimento sobre um determinado tema, definindo padrões para a organização de dados e metadados. Assim, uma ontologia oferece uma base estruturada para a interpretação e uso do conhecimento, promovendo clareza e consistência nas definições dentro de um domínio.

Existem diretrizes e boas práticas para se construir uma ontologia - essa ideia é resumida no conceito de *ontology design patterns* (padrões de design para ontologias). Por exemplo, a BFO (*Basic Formal Ontology*)[152] é uma ontologia de alto nível que fornece uma estrutura para a construção de ontologias específicas de domínio, organizando entidades em categorias gerais como objetos, processos e ações, para facilitar a interoperabilidade semântica entre sistemas. Enquanto a BFO pode ser considerada uma ontologia genérica, a SIO (*Semanticscience Integrated Ontology*) [153] é uma ontologia mais específica para a área científica, visando a integração e a interoperabilidade de dados científicos, com foco em representar conceitos científicos em áreas como biologia, medicina e química. Também existem outras ontologias de alto nível para domínios mais específicos como DOLCE (*Descriptive Ontology for Linguistic and Cognitive Engineering*) [154] e EMMO (*Elementary Multiperspective Material Ontology*) [155].

As tecnologias de implementação de ontologias envolvem ferramentas e linguagens específicas que permitem a modelagem, formalização e uso de conhecimento em domínios diversos. Entre as principais estão as linguagens OWL (*Web Ontology Language*) [156] e RDF (*Resource Description Framework*) [157], que possibilitam a representação semântica de dados, favorecendo a interoperabilidade entre sistemas. Essas representações podem ser serializadas por formatos como XML, Turtle [24] e JSON-LD[25] Ferramentas como Protégé [158], uma plataforma amplamente utilizada para edição de ontologias, e *reasoners* como Owlready [159] ou Pellet[160], que permitem a inferência lógica a partir das definições ontológicas, também são fundamentais no processo de construção e implementação de ontologias. Essas tecnologias são essenciais em áreas como Web Semântica, inteligência artificial e sistemas de informação, promovendo a organização e reutilização eficiente do conhecimento.

A ontologia se difere do FAIR no sentido que este é uma diretriz, enquanto aquela, quando estabelecida, configura-se um padrão de modelagem do conhecimento. Ter uma

[24]Turtle é uma linguagem de serialização compacta e legível para humanos usada para representar dados RDF, utilizando uma sintaxe simples baseada em prefixos e tripletas (sujeito, predicado, objeto).

[25]JSON-LD (*JavaScript Object Notation for Linked Data*) é uma maneira de representar dados estruturados de forma legível tanto para humanos quanto para máquinas, com foco em dados interligados (*Linked Data*). É uma mapeamento do RDF para o formato JSON.

ontologia bem definida pode, portanto, permitir que os princípios FAIR sejam atingidos. Vale lembrar também, que a própria construção de ontologias acumulou regras e boas práticas intrínsecas a área e que não correspondem as diretrizes do FAIR em seu objetivo (o de construir boas ontologias).

Estabelecendo esse conjunto de regras ao construir uma relação hierárquica sobre as informações e conhecimentos de um domínio, a ontologia pode servir a dois caráteres do FAIR: um relacionado a reprodutibilidade científica e a proveniência de uma informação; e outro relacionado a capacidade de se utilizar ferramentas programáticas para operar sobre esses dados, uma vez que os padrões sejam mantidos entre diferentes autores e plataformas. É importante notar que, contudo, a interoperabilidade dos dados por padrões específicos, por exemplo, o formato dos arquivos de dados implementados em um sistema computacional, não constituem a hierarquia e relacionamento do conhecimento em si (a ontologia), mas uma definição de boa prática a ser adotada.

No caso da Informática de Materiais, diversas ontologias foram propostas. A Materials Ontology é uma ontologia que estabelece os conceitos de substância, propriedade, ambiente e processo, tendo sido desenvolvida para a troca de dados entre bancos de dados heterogêneos de materiais, particularmente para propriedades térmicas [161]. Sua implementação pode ser via OWL com serialização em XML ou alguma versão específica para materiais como MatML [162] e MatDB [163]. A MatOnto [164] é uma antologia baseada em DOLCE com objetivo de acelerar a descoberta de novos materiais cristalinos através do compartilhamento de dados experimentais. A MDS-Onto (*Materials Data Science Ontology*) é uma proposta baseada em BFO com pacote em Python (FAIRmaterials) para criação de ontologias no domínio de Ciência de Dados de Materiais [165]. Neste trabalho os autores discutem duas implementações para os subdomínios de raio-x por radiação síncrotron e de materiais fotovoltaicos. A NanoMine Ontology é baseada em SIO aplicada no subdomínio de nanomateriais poliméricos [166]. Não se limitando as pesquisas em materiais em sua escala micro, nano, e atômica, a BMP (Ontology Building Material Performance) é uma ontologia para especificar materiais macroscópicos usados na construção civil e arquitetura, no contexto de aumentar a integração entre a modelagem informacional de materiais (*Material Information Modelling*, MIM) e a simulação de performance em construções (*Building Performance Simulation*, BPS) [167]. A MDO (*Materials Design Ontology*) é uma ontologia que prevê suporte ao padrão OPTIMADE de APIs, visando melhorar o acesso e interoperabilidade entre bases de dados de materiais computacionais [168, 169].

Por fim, chamamos a atenção para o fato de que a definição de ontologias exerce papel importante não só para a construção de bases de dados FAIR, mas também para validação de novas bases de dados recuperadas por mineração de texto. Por se basearem em engenharia de conhecimento de especialistas, as ontologias fornecem uma série de regras lógicas para validar os dados não estruturados extraídos ou recuperados de fontes textuais, permitindo que estes dados sejam restruturados para formatos mais amigáveis para máquinas - seja para facilitar o acesso (bases de dados), seja como dados de treinamento para modelos de aprendizado de máquina.

3.2.4 Mineração de Dados

Mineração de Dados é o processo de descobrir informação útil (padrões e relações) em grandes volumes de dados [170]. Uma característica importante desse processo é que ele é assistido tanto por técnicas computacionais relacionadas as tecnologias de armazenamento de informação, como banco de dados e recuperação de informação tradicional, quanto por técnicas computacionais de raspagem de dados da web (do inglês, *web scraping*) e de extração de informação. Ferramentas tidas como de inteligência artificial podem fazer parte dessas diferentes técnicas citadas.

A primeira etapa de mineração de dados envolve a coleta dos dados. Estes dados podem estar disponíveis e já estruturados em *datasets* ou bases de dados especificas, ou estarem dispersos, por exemplo, pela internet ou pela literatura científica. Independente da fonte, quanto mais estruturado o dado estiver, mais fácil será para ele ser extraído por uma máquina. Portanto, dados estruturados em ontologias facilitam a recuperação da informação através da estrutura de conhecimento embutida na sua construção; e uma fonte que provê suporte a APIs com requisições específicas torna-se uma peça essencial para definir o grau de automatização que a tarefa de coleta permite.

Por outro lado, grandes quantidade de dados estão em fontes que não provem esse suporte, por exemplo, a própria internet como um todo é uma base dados que não provê, necessariamente, acesso programático a todas as informações individuais via API, nem estão organizados em ontologias bem estruturadas e apontadas para um domínio de aplicação. Neste caso, são aplicadas técnicas de *web scraping*, nas quais *scripts* automáticos copiam dados brutos ou pré-processados de páginas web. Por trás desses *scripts* de *web scraping* podem existir ou não *scripts* de rastreamento de rede (também conhecidos como *web crawler* e *spiderbot*). Um *web crawler* processa uma fila de URLs adicionando a essa lista qualquer nova URL encontrada no recurso vasculhado atual (a URL atual). Isso é útil para construir uma rede de como os recursos (páginas web) estão interligados. Do ponto de vista do *web scraping*, o *web crawler* ajuda a aumentar o espaço de busca da informação que está sendo raspada. Em ambos os casos, o *script* de computador pode emular a ação de uma pessoa real acessando o recurso.

Entretanto, seja por *web scraping* ou mesmo com APIs específicas, alguns dados podem vir em formatos legíveis somente por humanos, como textos ou imagens. Tais técnicas são podem ser entendidas como técnicas de Recuperação de Informação (do inglês, *Information Retrieval*, IR), na medida que retornam uma informação que estava espalhada em diferentes documentos (como arquivos de texto puro e arquivos de hipermídia) sem gerar alteração nesta informação. Para que dados legíveis somente por humanos sejam entendidos pela máquina são necessárias técnicas de inteligência artificial[26]. Elas tem um papel central na mineração de dados pois são elas as principais responsáveis por extrair a informação quando os dados só podem ser lidos por humanos e quando os dados tem volume e dimensões

[26]Recuperação de Informação também pode usar IA, por exemplo, numa consulta de busca semântica, onde a informação a ser recuperada é escrita em linguagem natural, devendo o computador interpretar a quais dados nos arquivos pre-existentes aquela busca corresponde.

demais para serem totalmente analisados por humanos. Em termos gerais, as técnicas de IA voltadas à Extração de Informação (do inglês, *Information Extraction*, IE) de um dado não estruturado (texto, imagem, áudio) transformando-a em um dado estruturado são estudadas pelo Processamento de Linguagem Natural (NLP, do inglês, *Natural Language Processing*) para as conversões de texto, a Visão Computacional para a conversões de imagens e vídeos, e o Processamento de Áudio para as conversões de áudio contendo fala humana ou outro evento sonoro de interesse; já as técnicas relacionadas à extração de informação de dados estruturados ou não, focando na questão da quantidade (de dimensão e volume) são objeto de estudo da Ciência de Dados.

A seguir discutiremos duas situações relevantes da aplicação de Mineração de Dados no contexto de Ciência de Materiais.

3.2.4.1 Screening de Materiais

No Informática de Materiais, a tarefa de *screening* refere-se à coleta e a seleção de compostos de bases de dados. A coleta pode ser assistida por uma API que provê acesso programático à base de dados. Já a seleção se baseia em filtros que operam, normalmente, sobre determinadas propriedades reportadas dos compostos. Por exemplo, pode-se filtrar materiais pelo tipo de sistema cristalino, pelo bandgap, pela quantidade de átomos, e assim por diante. Filtros são realizados em cascata para que os materiais selecionados atendam vários critérios. Em geral, o *screening* ocorre sobre bases de dados estruturadas, normalmente em estruturas tabulares (tuplas de entrada e atributos).

Esses processos objetivam tanto a seleção de um único material quanto a construção de *datasets* com propriedades de interesse. Esses *datasets* podem ser usados na sequência para análises descritivas ou preditivas. Por exemplo, *Acosta et al.* realizaram *screening* no ICSD com filtros para identificar materiais topológicos e com Efeito Rashba [171]. Já *Zhuo et al.* agregaram dados de bases de dados experimentais e do Materials Project para treinar algoritmos de predição do bandgap dos materiais [172], o que se torna uma ferramenta de *screening* (filtro preditivo) para definir se um material é isolante ou não.

O *screening* é, em geral, relativamente simples de ser implementado quando as bases de dados de origem tem ontologias bem definidas (informação estruturada) e provem APIs de acesso.

3.2.4.2 Mineração de Texto

Texto é um formato legível por humanos, onde a informação está não estruturada e é altamente heterogênea [173]. Por isso, técnicas de mineração de texto precisam transformar a informação textual em um formato estruturado que possa ser mais facilmente manipulado por máquinas.

A fonte dos textos pode ser a mais variada. Por exemplo, dados de redes sociais podem ser raspados ou acessados via APIs para monitorar reações de usuários a determinadas

campanhas de marketing. No contexto de Ciência de Dados de Materiais, os textos utilizados são a literatura científica.

O primeiro passo da mineração de texto neste cenário é a obtenção dos artigos. Isso pode ser feito via APIs que as editoras disponibilizam. Em geral os texto podem vir em formato PDFs (*Portable Document Format*) ou XML (*Extensible Markup Language*). Do ponto de vista programático, XML é um formato mais fácil de se trabalhar, entretanto em ambos os casos há dificuldades em relação à padronização das estruturas internas do arquivo pelas diferentes editoras como seções com nomes diferentes e/ou inexistentes de uma revista para a outra. Há também questões relacionadas aos conteúdos não diretamente verbais como tabelas gráficos, imagens e fórmulas. Estes elementos geram mais dificuldades para a extração de informações: por exemplo, imagens precisam ser processadas por técnicas de IA adicionais relacionadas à visão computacional para terem suas informações relevantes extraídas.

Existem vários exemplos de aplicações de NLP e mineração de texto para a área de materiais. Tshitoyan *et al.* mineraram texto de $3,3$ milhões de resumo de artigos publicados entre 1922 e 2018, estabelecendo um vocabulário de 500 mil palavras que posteriormente foi usado por técnicas de aprendizado de máquina para criar um sistema capaz de prever a aplicação de um material anos antes daquele material de fato ser aplicado naquele domínio [174]. Kononova *et al.* construíram um *dataset* sobre síntese de materiais inorgânicos analisando mais de 53 mil parágrafos de artigos da literatura usando técnicas de *web scrapping* [175]. *Court et al.* demonstraram como utilizar técnicas de extração de entidades de grande volume de texto para construir *datasets* com propriedades de interesse para treinamento de sistema de predição de propriedades supercondutoras [176] e magnetocalóricas [177]. De Paula estabeleceu um fluxo de processamento que pode partir de arquivos em formato `pdf`, misturando técnicas de extração de entidades nomeadas (NER, *Named Entity Recognition*) baseada em regras e classificação utilizando aprendizado de máquina [178]. Bandeira *el al.* utilizaram técnicas de NER por regras e por modelos generativos para extrair eficiência faradaica para redução de CO_2 de mais de 7 mil artigos [179].

O estado da arte das técnicas de NLP combinadas com IA são os modelos fundacionais de texto [180]. Esses modelos amplamente conhecidos como *Large Language Models* (LLMs) são sistemas de aprendizado de máquina treinados com grande quantidades de texto para realizar, de maneira geral, a tarefa de predição de palavras. Entretanto, à medida que a escala desses sistema aumenta, propriedades emergentes surgem, de maneira que tais modelos podem ser aplicados para tarefas diferentes daquelas para as quais eles foram treinados [181, 182]. Exemplos de tarefas que podem ser realizadas para as quais os modelos não foram treinadas inclui NER, classificação de texto, geração de código de computador (geração de linguagem formal), tradução de texto, extração de informação, entre outras.

A primeira técnica precursora dos LLMs foi o word2vec [183]. Este modelo encontra uma representação vetorial para palavras contidas em um texto[27], analisando as estatís-

[27]Na memória de um computador, as palavras são representadas em alto nível como estruturas denomi-

ticas de ocorrência e coocorrência das mesmas. Estas representações vetoriais são mais fáceis de serem processadas por outros algoritmos computacionais, especialmente aqueles relacionados ao aprendizado de máquina. Com isso é possível extrair relações em grande quantidade de texto, por exemplo, calculando o cosseno entre a representação vetorial de duas palavras diferentes, isto é, medindo o grau de similaridade entre elas sem utilizar um conhecimento humano prévio (apenas a representação vetorial extraída pelo modelo word2vec). É exatamente esta a técnica principal utilizada por Tshitoyan *et al.* [174] comentada acima. Por outro lado, estas técnicas possuem limitações quanto a semânticas diferentes que uma mesma palavra pode transmitir em contextos diferentes. Para superar essas limitações, modelos de aprendizado profundo utilizando técnicas de redes neurais artificial foram desenvolvidos. A principal técnica é chamada de *atenção*, que envolve estabelecer, durante o processo de transformação das palavras em vetores, *tokens* adicionais para contextos diferentes [184]. Dessa ideia fundamental surgem modelos mais robustos, como o BERT (*Pre-training of Deep Bidirectional Transformers*) [185], o BART [186], posteriormente seguidos de uma família de modelos GPT [187] (*Generative Pre-trained Transformer*). As técnicas usando atenção podem ser dividas quanto a forma de treinamento desses modelos, que se divide em somente codificador (*encoder-only*) como o BERT, codificador-decodificador (*encoder-decoder*) como o BART ou somente decodificador (*decoder-only*) como o GPT [187].

Modelos do tipo BERT conseguem gerar boas representações contextuais das palavras que posteriormente são usadas para treinar tarefas específicas (conhecido como *downstream tasks*) ou ajustados para corpus[28] de domínios específicos (conhecido como *fine-tuning*). É comum essas técnicas serem encadeadas. Um exemplo na área de materiais é o MatSciBERT, onde um BERT pré-treinado passou por um *fine-tuning* para domínio científico (SciBERT) [188], seguido por outro *fine-tuning* para domínio de materiais [189]. O MatSciBERT demonstrou capacidades de realizar NER em um *dataset* sobre SOFC (célula de combustível de óxido sólido, do inglês, *solid oxide fuel cell*), classificação de relações entre entidades em *dataset* de síntese de materiais e classificação binária de resumos em se tratando de materiais vítreos ou não, com performances ligeiramente superiores aos modelos BERT e SciBERT [189].

Porém os avanços mais recentes sugerem que modelos generativos capazes de realizar tanto tarefas de Entendimento de Linguagem Natural (NLU, *Natural Language Understanding*) quanto tarefas de Geração de Linguagem Natural (NLG, *Natural Language Generation*) conseguem, quando atingem escalas de bilhões de parâmetros, executar diferentes *downstream tasks* sem serem explicitamente programados para isso. Essas propriedades emergentes podem ser controladas pela sequência inicial de palavras fornecida ao modelo. A ideia geral é, de maneira simplificada, que, ao predizer as probabilidade de uma próxima

nadas *strings*. Tais estruturas não permitem, de maneira direta, a utilização de métodos numéricos. Dessa forma, o objetivo da representação vetorial é representar uma *string* como um vetor de números reais.

[28]Em NLP, corpus é um conjunto de textos que constitui um dataset sobre um domínio (assunto dos textos). O termo também é usado para designar conjuntos de documentos, mais especificamente em tarefas de recuperação de informação.

palavra ocorrer dado uma sequência de palavras de entrada, o modelo acaba por gerar texto que corresponde a realizar da tarefa pedida na sequência original. Diferentes exemplos são fornecidos na literatura [182, 190, 191]. A escolha adequada das palavras de entrada que geram os resultados esperados é chamada de *engenharia de prompt* [192] e estão relacionadas a capacidades desses modelos de apresentar aprendizado em contexto [181]. Por exemplo, Polak e Morgan demonstraram um simples protocolo de engenharia de *prompt* para utilizar um LLM da família GPT (chatGPT) para realizar NER em textos passados como entrada, obtendo tuplas de material, valor e unidade de propriedades de interesse pedidas (também no texto de entrada) [192]. Contudo, as capacidades desses modelos não se limitam à mineração de texto: utilizando a versão GPT4, um time de pesquisadores da Microsoft conseguiu demonstrar capacidade desses modelos de, inclusive, performar partes de contas DFT como geração de arquivos de estruturas e predição de energia final [191].

Por ser exatamente o estado da arte nesse tipo de tecnologia, ainda existem discussões sobre as reais capacidades desses modelos. Por outro lado, a versatilidade desses sistemas permite a sua exploração por áreas diferentes, como a Ciência de Materiais. Como demonstramos com alguns exemplos, é provável que essa ferramenta torne-se-a elemento fundamental para a Ciência de Dados de Materiais. Porém, não será a única produzida a partir de técnicas de IA a ser aproveitada por essa área: treinamento de modelos de IA para predição de propriedades de interesse tem grande potencial de encurtar o tempo para a descoberta de materiais. Dessa forma, a seguir, fazemos uma breve sistematização de diferentes técnicas de IA sendo diretamente aplicadas na área de materiais.

3.3 Inteligência Artificial em Ciência dos Materiais

3.3.1 Representação de Dados

Do ponto de vista da Ciência de Materiais, a escolha da representação da entrada de uma algoritmo de ML é uma etapa fundamental. Reduzindo o problema a representação de um material, o espaço de *features* possíveis pode ser divido em **features composicionais** e **features estruturais**. Atributos composicionais se referem a propriedades dos átomos constituintes do material. Podemos representar cada átomo do material por um conjunto de atributos atômicos, concatenando essa representação em um único vetor no qual cada dimensão é uma propriedade de um átomo específico (em geral, limitado a átomos únicos). Por exemplo, podemos representar o CaF_2 como $(20, 1, 1, 9, 3.98, 2)$ onde as três primeiras dimensões do vetor representam, respectivamente, o número atômico, a eletronegatividade e a quantidade de átomos de cálcio, enquanto as três últimas representam as propriedades na mesma ordem, só que do flúor. Podemos também utilizar valores estatísticos como média aritmética, geométrica, harmônica, ponderada, máximo ou mínimo das propriedades dos átomos para permitir utilizar, em um mesmo *dataset*, materiais com qualquer número diferente de átomos únicos (binários, ternários, etc.).

Diferentes trabalhos utilizaram essas técnicas para treinar algoritmos de ML para re-

alizar tarefas, principalmente, de classificação e de regressão. Por exemplo, Wang *et al.* treinaram algoritmos de ML para classificar perovskitas duplas A_2BBX_6 entre isolante, semicondutor ou metal usando 14 *features* composicionais, envolvendo propriedade atômicas diretas e calculadas (diferença entre o sitio A e X) [193]. Acosta *et al.* resolveram um problema de classificação de ordem magnética em materiais 2D utilizando como *feature* de entrada, a média aritmética entre as propriedades atômicas que incluíam número atômico, eletronegatividade de Pauling, grupo da tabela periódica, número de valência e raios atômicos [194]. Já Zhuo *et al.* utilizaram 136 *features* atômicas para treinar um algoritmo de regressão capaz de predizer valores para o bandgaps de sólidos inorgânicos [172], o que pode ocasionar um tipo diferente de *overfitting* relacionado à *maldição da dimensionalidade* - isto é, o tamanho do espaço combinacional das *features* ser muito maior do que a informação necessária para generalização oriunda da amostragem de exemplos fornecidos durante o treinamento do modelo; ou, por outro ponto de vista, o modelo tem mais parâmetros do que estatística relacionada ao problema, ocorrendo um sobreajuste similar a ajustar um polinômio de grau muito grande para uma curva qualquer.

Apesar de serem técnicas com grande qualidade preditiva, a utilização de representações por *features* composicionais apresenta uma desvantagem importante: ela não é capaz de endereçar materiais com a mesma composição porém com estruturas diferentes.

As *features* estruturais devem demonstrar completude, invariância, suavidade e baixa complexidade. Enquanto a completude garante que estruturas não equivalentes sejam representadas para atributos distintos, a invariância requer que estruturas equivalentes sejam representadas pelos mesmos atributos de entrada. A suavidade implica que deformações contínuas de uma estrutura devem ser mapeadas para uma deformação suave das características representadas. Por fim, baixa complexidade refere-se a complexidade computacional de representar o material, que deve exigir o menor esforço possível em tempo e memória.

A biblioteca DScribe [195] implementa em Python uma coleção de métodos para geração de descritores estruturais, como a Matriz de Coulomb [196], a Matriz Seno [197], a Matriz de soma de Ewald [197], o ACSF (*Atom-centered Symmetry Functions*) [198], o SOAP (*Smooth Overlap of Atomic Positions*) [199], o MTBR (*Local Many-body Tensor Representation*) [200] e o descritor de Valle-Oganov [201]. Cada um desses descritores tem características próprias, por exemplo, a Matriz de Coulomb, apesar de simples, consegue capturar elementos que permitem comparar moléculas enquanto que a Matriz Soma de Ewald é uma extensão da Matriz de Coulomb para sistemas periódicos assim como a Matriz Seno.

O estado da arte, contudo, explora outras formas de representação mais sofisticadas. Por exemplo, *voxels*[29] foram usados para codificar a densidade eletrônica, com essa representação sendo usada para entrada em redes neurais convolucionais [202]. Outro exemplo é a utilização de grafos para representar a relação espacial entre os átomos e um vetor de *features* atômicas para representar a espécie química. Exemplos de algoritmos de ML

[29]Um voxel é a menor unidade volumétrica de uma representação tridimensional, análogo a um pixel em 2D.

implementados com essa abordagem são os modelos da família ALIGNN (*Atomistic Line Graph Neural Networks*) [203], M3GNET[204] e MACE [205], todos exemplos de redes neurais artificiais.

3.3.2 Interpretabilidade e Explicabilidade

Além da questão de representar informação estrutural e composicional, a escolha de *features* é um momento oportuno para construir modelos de aprendizado de máquina que sejam interpretáveis. Isso é especialmente útil quando queremos extrair entendimento físico e químico das predições ou descrições que fazemos com esses algoritmos.

Existem diferentes técnicas de interpretação de *features* que são independentes do domínio de aplicação. Por exemplo, alguns algoritmos de ML fornecem uma pontuação de importância para cada *feature*. Isso é fácil de entender no caso da Regressão Linear: se as *features* estão na mesma escala, então o valor dos coeficientes associados a cada uma das *features* representa a intensidade com que a variação daquela *feature* afeta a variável dependente de saída. Podemos usar essa informação, inclusive, para reduzir a dimensão do problema eliminando *features* de baixa importância. Nesse sentido, espera-se que um modelo com menor número de *features* seja mais facilmente interpretável. Além disso, ao atestar a importância de uma *feature* temos um indicativo do tipo de variável que importa para a solução do problema.

O método de regressão LASSO (*Least Absolute Shrinkage and Selection Operator*), adiciona penalização a função custo, o que faz com que alguns dos coeficientes do modelo se aproximem de zero, reduzindo a dimensão ao mesmo tempo em que aprende a realizar as predições. Entretanto essa metodologia falha quando os espaço de *features* é muito grande ou há muitas *features* correlacionadas [206].

Por outro lado, se escolhemos *features* que representam entidades em si interpretáveis, podemos interpretar o modelo reduzido de acordo com sua capacidade de predição. Se pegarmos o exemplo da Regressão Linear aplicada sobre dados de corrente elétrica I e tensão elétrica V, o modelo fenomenológico obtido é exatamente a Lei de Ohm ($V = RI$), com o coeficiente angular representando a resistência elétrica R - uma propriedade física [207].

Isso é factível uma vez que o modelo usado (Regressão Linear) é paramétrico, assumindo uma equação (função linear) como relação entre as *features* de entrada e a variável de saída. Entretanto, por vezes são utilizados algorítimos não-paramétricos (como Árvores de Decisão e algoritmos de *ensemble* como RandomForest e GradientBoosting) ou cujos parâmetros não geram uma interpretação analítica (como RNAs), de forma que o modelo preditivo não é interpretável por uma equação.

Sistemas de Regressão Simbólica tentam descobrir qual é a expressão matemática que relaciona as *features* a tarefa pretendida (regressão ou classificação). Em geral, são estabelecidos conjuntos de operações que podem ser feitas entre as *features* e alguns coeficientes arbitrários (por exemplo, soma, subtração, multiplicação e divisão). Em seguida, uma série

de tentativas são feitas para se encontrar uma expressão cuja regressão retorne o menor erro.

Uma abordagem para se fazer Regressão Simbólica é através do uso de programação genética, como implementada pelo gplearn [208]. Esse tipo de abordagem utiliza elementos de evolução biológica para modificar populações de soluções representadas por diferentes combinações entre operações, coeficientes e *features*. Por outro lado, em se tratando de fenômenos físicos, Udrescu e Tegmark propuseram um modelo em cascata que utiliza análise dimensional e operações de simetria para construir equações fisicamente razoáveis [209]. Para a área de materiais, Ouyang *et. al* propuseram o SISSO (*Sure Independence Screening and Sparsifying Operator*) um regressor simbólico que realiza redução de dimensionalidade durante o processo de descoberta da relação entre as *features* [206]. Mais recentemente, uma implementação em Python do SISSO é discutida em [210]. Todos estes modelos tem o diferencial de forneceram uma equação no final do processo de aprendizado.

A interpretabilidade de métodos de regressão simbólica são inerentes ao modelo, enquanto métodos baseado na importância de *features* são ditos *post-hoc*. Métodos *post-hoc* são independentes do modelo de aprendizado de máquina usado, envolvendo importância de *features* medida através de permutação, como o método PFI (*Permutation Feature Importance* [211], ou teoria do jogos, como o SHAP (*SHapley Additive exPlanations*) [212]. Outra abordagem é através dos gráficos ICE (*Individual Conditional Expectation*) que mostram como cada predição por instância do *dataset* muda quando uma determinada *feature* muda [213]. Por outro lado os gráficos PDP (*Partial Dependence Plots*) sintetizam o efeito médio de uma variável sobre as predições do modelo [214], enquanto os gráficos ICE detalham esse efeito em nível individual, possibilitando a identificação de heterogeneidades e interações que seriam obscurecidas pela média global. Também é possível criar um novo modelo de ML interpretável para imitar o comportamento preditivo de um modelo não interpretável usando o método LIME (*Local interpretable model-agnostic explanations*) [215].

Em Ciência de Dados de Materiais Lee e Miyamoto utilizaram medidas estatísticas de correlação, PFI e SHAP para interpretar modelos de predição de bandgap, ajudando a eliminar *features* menos importantes [216]. Já Masuda e Tanabe exploraram o mesmo problema utilizando PFI, PDP, ICE, e ALE (*Accumulated local effect plot*) para identificar a influência das *features* atômicas na predição do bandgap de dados computacionais (Materials Project) e experimentais [217]. Enquanto estes são exemplos de métodos *post-hoc* agnósticos ao modelo de predição, Ogoshi *et al.* utilizaram o VAX (*multiVariate dAta eXplanation*) para identificar regras que determinariam se um material tem bandgap direto ou indireto [218].

Todos esses métodos partem de *features* atômicas ou assinaturas estruturais que podem ser representadas por vetores de *features*. O uso de modelos de representação mais complexos como *voxels* e grafos, ainda constitui um desafio para a explicabilidade e a interpretabilidade. Mesmo as representações vetoriais ainda são pouco exploradas para XAI dentro Ciência de Dados de Materiais, constituindo portanto, uma frente de avanço promissora para trabalhos futuros.

3.3.3 Modelos Substitutos

Modelos substitutos (conhecidos em inglês como *surrogate models*) são uma classe de modelos computacionais usados para aproximar ou substituir simulações ou experimentos complexos e caros, proporcionando uma alternativa mais econômica e rápida.

Métodos quânticos como Hartree-Fock e DFT tem complexidade de tempo polinomial $\mathcal{O}(N_e^4)$ sobre o número de elétrons N_e [219]. Na prática, o custo computacional também depende do grau de complexidade do funcional de troca e correlação assumido, já que, mesmo sendo cálculos cujo tempo assintótico de execução não muda com o número de elétrons, as constantes de tempo específicas dependem de cada implementação. Métodos heurísticos implementados pelos modelos substitutos, por outro lado, podem ser muito mais rápidos de serem calculados, alguns apresentando tempo $\mathcal{O}(1)$ independente do tamanho da entrada. A vantagem, portanto, de se utilizar um método heurístico é acelerar os testes.

Em geral um modelo substituto pode ser usado: como um elemento da simulação DFT, por exemplo, ajudando no cálculo da estrutura eletrônica ou sendo treinado com esta para acelerar a determinação de uma determinada propriedade; como um potencial interatômico a partir do treino com cálculos de AIMD; como um preditor de propriedades baseado em descrições *ad hoc*, como *features* composicionais. Abaixo fazemos uma distinção entre essas três perspectivas de aplicação dos modelos substitutos:

- Sistemas Preditivos: Algoritmos de ML baseados em *features* que são facilmente obtidas (por exemplo, *features* atômicas) são usados para predizer uma propriedade final que é de difícil obtenção ou que não se conhece uma expressão de relação entre ela e as variáveis de entrada. Por exemplo, predição de bandgap experimental a partir de *features* atômicas [172].

- Sistemas Corretivos: Algoritmos de ML cuja pelo menos uma das *features* de entrada é da mesma natureza da variável de saída. Por exemplo, treinar um modelo com bandgap PBE-GGA para prever o bandgap HSE [220] ou bandgap GW [216].

- Potencial Interatômico: Algoritmos de ML aplicados para aprender potenciais interatômicos (modelos denominados MLIP, do inglês, *Machine Learning interatomic Potentials*), auxiliando em contas de dinâmica molecular [221, 222].

3.3.4 Modelos Generativos

Modelos de IA generativos são algoritmos capazes de aprender padrões em dados e criar novas amostras realistas que seguem essas distribuições, como textos, imagens ou sons. Para a Ciência de Dados de Materiais, em especial para o tema tratado neste trabalho, estamos interessados em gerar novos materiais.

Para o domínio computacional, "gerar um material" irá significar "gerar uma descrição" de um material. Por exemplo, queremos gerar um arquivo CIF (*Crystallographic Information File*) que contém informações sobre as coordenadas de espécies químicas em uma estrutura cristalina.

Logo, ao tentarmos descobrir estes novos materiais, um problema central que estamos tentando resolver é a predição de estruturas cristalinas (do inglês, *Crystal Structure Prediction*, CSP) estáveis. Este problema envolve um espaço de busca composicional (espécies químicas) e estrutural (configuração espacial dos átomos), e uma forma de atestar a estabilidade através do ranqueamento das estruturas pelas suas energias [223].

O conceito de espaço composicional pode ser entendido como a quantidade C de combinações permitidas por n elementos da tabela periódica, agrupados em k, isto é:

$$C_k^n = \frac{n!}{k!(n-k)!}. \tag{3.15}$$

Considerando $n = 103$ espécies químicas diferentes, para compostos binários ($k = 2$), temos 5253 possíveis sistemas; para ternários ($k = 3$), 176851; e para quaternários ($k = 4$), 4421275 [224]. Estes são sistemas e não materiais de fato: para cada sistema temos diferentes estruturas, quantidades de átomos e estados de oxidação possíveis para cada composto contendo as espécies químicas daquele sistema. Considerando as regras adicionais de neutralidade de carga e o balanço de eletronegatividade, o espaço composicional fica restrito a 14721 compostos binários, 32157899 ternários e 32381953858 quaternários [224].

O espaço de busca total, contudo, é ainda maior, já que uma mesma composição química pode apresentar uma estrutura diferente. Considerando arranjos de N átomos com d graus de liberdade (para átomos descorrelacionados $d = 3N+3$), o número de estruturas possíveis é $E \sim e^{ad}$, onde a é uma constante específica para o sistema, o que é aproximadamente $E \sim 10^N$ [223]. O espaço de busca total é, portanto, algo como:

$$\Omega = C \times E \sim \frac{n!}{k!(n-k)!} 10^N, \tag{3.16}$$

com $k \leq N$.

Isso significa que o espaço de busca tem complexidade exponencial com o número de átomos, o que denotamos por $\mathcal{O}(l^N)$, onde l é um número natural qualquer (geralmente $l = 2$, mas estritamente na Eq. 3.16, $l = 10$). Isso significa que o prolema CSP é do tipo NP [78].

O problema CSP pode ser resolvido por um método de simulação física, como o DFT, ou por um modelo substituto, como os sistemas preditivos. Também podemos fazer soluções híbridas, por exemplo com um modelo de potencial interatômico para simular estruturas em diferentes temperaturas em uma simulação de dinâmica molecular. Nos referimos a essas soluções como módulos testadores, isto é, métodos que irão testar se a estrutura proposta é de fato estável.

O módulo testador, entretanto, é só uma parte de um modelo generativo: ainda é necessário gerar o material que será testado, ou seja, precisamos de um módulo gerador. Módulos geradores podem empregar diferentes técnicas. Por exemplo, é possível utilizar algoritmos genéticos para propor novas estruturas ou composições [225, 226].

Também podemos utilizar métodos baseados na amostragem do espaço latente encon-

trado por redes neurais *autoenconder* [227, 228, 229]. A ideia aqui e utilizar a RNA para encontrar uma nova representação do material (o espaço latente, normalmente com uma dimensionalidade menor que o espaço das variáveis de entrada, sendo compostos por combinação não linear das mesmas) e então amostrá-lo para gerar novos dados. Esses dados gerados no espaço latente são decodificados para o espaço de entrada na saída da rede *autoenconder*, gerando assim um possível novo material. Se a rede for treinada visando minimização de energia das estruturas, as representações aprendidas podem corresponder a estruturas estáveis.

Outra possibilidade é a utilização dos LLMs discutidos na Seção 3.2.4.2: podemos pedir para um modelo de texto gerar o arquivo CIF de uma nova estrutura que atenda determinadas propriedades [230]. Ainda há discussão de se fato o modelo consegue criar uma estrutura totalmente nova ou apenas "lembrar"de estruturas que foram utilizadas no treinamento, porém, é claro que os elementos não-determinísticos desses modelos permitem certo grau de geração. Por outro lado, a arquitetura da rede neural *transformer* que é utilizada nesses LLMs também pode ser adaptada para trabalhos específicos com materiais [231].

Do domínio da geração de imagens, redes GAN (*Generative Adversarial Networks*) e redes de difusão também podem ser adaptadas para o domínio de materiais [232, 233, 234]. Chamamos a atenção para uma mudança interessante neste cenário: redes UniMat [233] são propostas de representação universal de materiais para modelos de difusão, elaboradas por pesquisadores do Google, enquanto a MatterGen é uma rede de difusão utilizada para geração de materiais proposta por pesquisadores da Microsoft [234]. O Google também é responsável pelo GNoME, uma rede neural profunda baseada em representação por grafos para a geração de novos compostos [235]. Vemos, portanto, uma tendência de exploração das tecnologias de IA para a descoberta de materiais sendo realizada por empresas privadas.

Para encerrar esse panorama geral sobre modelos generativos, gostaríamos de lembrar que modelos de predição também podem ser usados como módulo gerador. Dizemos que nesse caos temos um sistema generativo orientado por recomendação. O sistema de predição recomenda uma alteração da estrutura, como a composição química ou uma nova distância entre os átomos da rede, criando assim um novo material. Além disso, o módulo gerador pode ser tão prosaico quanto a substituição química dada por heurísticas e intuições já presentes na própria representação da tabela periódica.

3.3.5 Laboratórios Autônomos

Laboratórios autônomos são outro exemplo de aplicação de IA e aprendizado de máquina na área de materiais. Aqui o objetivo é utilizar sistemas robóticos para automatizar etapas de experimentação em laboratórios. As técnicas podem incluir desde a síntese do material até sua caracterização.

Existem diferentes graus de autonomia que podemos endereçar com essas técnicas. A ideia geral, não é totalmente nova: desde que um equipamento possa tomar medidas de

maneira automática, estamos automatizando o laboratório. Isso é muito comum na maioria dos equipamentos utilizados para caracterização, como DRX com porta amostras giratório que pode ser programado para uma rotina de medidas em diferentes amostras. Normalmente, quanto mais complexa a técnica experimental, menos automatizada ela é, exigindo intervenção humana na configuração e tomada de medidas. A direção que o paradigma orientado à dados fornece é relacionada as diretrizes operacionais: os equipamento mais modernos precisam ser equipados com softwares que permitam interoperação e configuração programática. Isso vai desde o formato em que os dados são gerados e armazenados (arquivos `.csv` e `.json`, entre outros, que fornecem um padrão aberto) até a possibilidade de integrar os softwares de controle desses equipamentos por APIs. Existem soluções comerciais como o LabVIEW (*Laboratory Virtual Instrument Engineering Workbench*), uma plataforma de programação gráfica desenvolvida pela National Instruments, amplamente utilizada para aquisição de dados, automação de instrumentos e controle de experimentos em laboratórios [236]. Evidentemente essa necessidade de interoperação está inerentemente limitada a natureza técnica e comercial dos equipamentos.

Do ponto de vista da implementação de laboratórios autônomos, podemos distinguir o uso de IA em dois níveis hierárquicos: no primeiro nível, as técnicas de IA são usadas para criar sistemas preditivos, generativos e de recomendação, visando sugerir um determinado sistema a ser processado; no segundo as técnicas de informática e IA são usadas para automatizar os equipamentos, incluindo treinar sistemas robóticos para reconhecer e manipular objetos.

O desenvolvimento de sistemas robóticos por sua vez, pode ser entendido pela automatização (como braços robóticos em linhas de montagem) e pela autonomia (como robôs androides que interagem livremente com o ambiente). No primeiro caso, o design da tarefa é adaptado ao mecanismo robótico. No segundo, os processos são mantidos e o sistema robótico deve ser adaptável o suficiente para conseguir operar sobre eles. Essa flexibilidade por parte do processo ou do sistema robotizado pode variar de grau entre diferentes abordagens, umas com sistemas mais autônomos (tomam decisões de maneira independente) ou mais automatizados (reproduzem rotinas pré-estabelecidas).

Na área biológica a automatização de experimentos é comumente conhecida como *hight throughput screening* [128, 237], normalmente envolvendo o teste de milhares de fármacos por equipamentos dedicados ou braços robóticos adaptados com micropipetadores. Para área de materiais os termos laboratórios autonomos (do inglêsm *autonomous laboratory*) e laboratório autoguiado (do inglês *Self-driving laboratory*) são termos mais comuns. Como exemplos de aplicações reais desses sistemas podemos citar a síntese de pontos quânticos [101], teste de propriedades mecânicas [238, 239], descoberta de filmes finos [102] e de cristais inorgânicos [240].

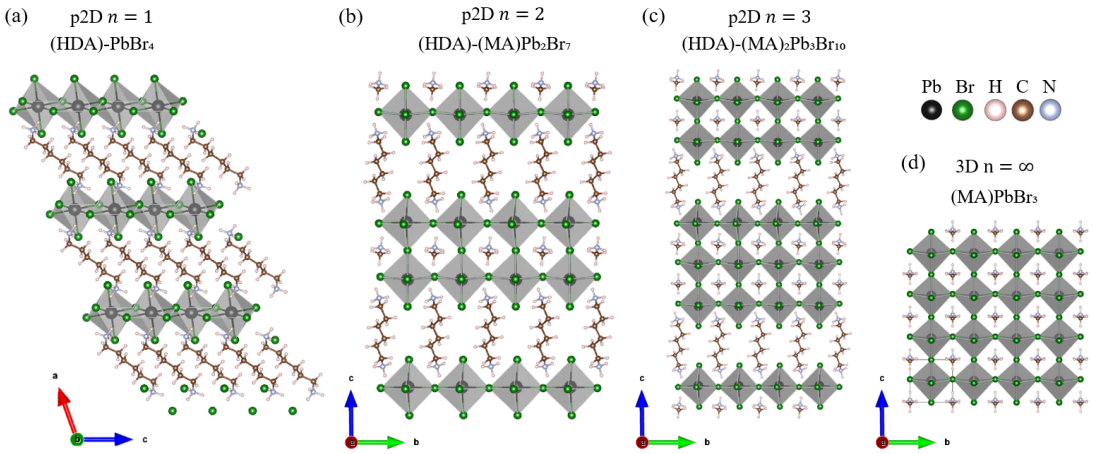
4 Estudo de Caso I: Perovskitas Híbridas pseudo-2D Dion-Jacobson

Perovskita é uma classe de materiais cristalinos com composição ABX_3, onde A e B são cátions de tamanho diferentes e X é um ânion, geralmente um haleto ou oxigênio. Os íons haletos formam vértices em octaedros centrados pelo cátion B e espaçados entre si pelos cátions A, com todos os vértices conectados em uma estrutura, geralmente, cúbica. A origem do nome está no mineral $CaTiO_3$, descoberto por Gustavo Rose em 1839 e denominado perovskita em homenagem ao mineralogista Lev Perovski [241].

As perovskitas híbridas orgânica-inorgânica (do inglês, *Hybrid Organic-Inorganic Perovskite*, HOIP) são perovskitas nas quais um dos sítios está ocupado por moléculas orgânicas [241, 242]. Nas HOIPs de haletos, $X = \{Cl^{1-}, Br^{1-}, I^{1-}\}$, e as moléculas orgânicas que ocupam o sítio A são, normalmente, metilamônio ($CH_3NH_3{}^{1+}$, sigla MA) ou formamidínio ($HC(NH_2)_2^{1+}$, sigla FA), enquanto o sítio $B = \{Pb^{2+}, Sn^{2+}, Ge^{2+}\}$ [241, 243]. Nas perovskitas de haletos inorgânicas, o sítio A é ocupado por elementos monovalentes grandes como metais alcalinos, comumente, césio Cs^{1+} [244].

Perovskitas 2D (Figura 4.1) são perovskitas híbridas nas quais são usadas moléculas orgânicas grandes nos sítios A, de forma a separar as camadas de octaedros. Essas moléculas são denominadas espaçadores orgânicos. Como a estrutura cristalina do material ainda reflete uma estrutura 3D, isto é, a periodicidade ocorre nas três dimensões do espaço, é preferível utilizar o termo pseudo-2D (p2D) [245], quasi-2D [246] ou pervoskitas em camadas (*layered perovskites*) [247, 248, 249, 250, 251] para designar essas perovskitas.

Figura 4.1: Representações dos cristais obtidos da relaxação por DFT. (a) $(HDA)PbBr_4$, (b) $(HDA)(MA)Pb_2Br_7$, (c) $(HDA)(MA)_2Pb_3Br_{10}$ e (d) $(MA)PbBr_3$.



Fonte: elaborado pelo autor.

Se as moléculas espaçadoras forem monovalentes a perovskita pseudo-2D será denominada de Ruddlesden–Popper (RP) enquanto se tais moléculas forem divalentes, as pervoskitas serão ditas de Dion–Jacobson (DJ) [248]. A perovskita é denominada ACI (do inglês, *alternating cations in the interlayer space*) quando há alternação entre as moléculas espaçadoras em uma mesma camada de separação dos octaedros [252]. Assim, as composições químicas dessas perovskitas podem ser representadas por $A'_2A_{n-1}B_nX_{3n+1}$ para RP, $A'A_{n-1}B_nX_{3n+1}$ para DJ, e $A'A_nB_nX_{3n+1}$ para ACI [253].

Na Figura 4.1 (a-c), a molécula espaçadora é a hexametilenodiamina (sigla HDA) divalente formando uma perovskita p2D do tipo Dion–Jacobson (a valência ocorre nos grupos amina em cada uma das extremidades da molécula). Para (b) e (c), além das moléculas espaçadoras de HDA, há moléculas de MA no sítio A interno à camada de octaedros. Já a pervoskita 3D em (d) não possui moléculas espaçadoras, com todos os sítios A ocupados apenas por metilamônio.

Há na literatura uma discussão sobre a dimensionalidade de uma perovskita quanto a sua estrutura eletrônica e estrutura morfológica [248, 254]. Cristais 2D como grafeno e fosforeno possuem uma única camada atômica de fato: a repetição estrutural acontece em dois eixos, x e y, enquanto no eixo z o material é cercado de vácuo. Esses materiais 2D constituem camadas de materiais *bulk* a grafite e o fósforo negro, respectivamente. Uma perovskita hibrida em camadas do tipo RP, pode ser comparada à grafite natural na qual a interação entre as camadas de grafeno ocorre devido a forças de van der Waals: numa perovskita RP há duas moléculas por sítio A, logo a interação entre essas moléculas na extremidade oposta à camada de octaedros é do mesmo tipo. Mesmo uma perovskita DJ, onde as moléculas divalentes realizam ligações químicas nas duas extremidades, o processo de esfoliação comum para a obtenção de cristais 2D a partir de materiais *bulk* em camadas, pode ser efetivamente usado para criar monocamadas atômicas de octaedros [248]. Este tipo de dimensionalidade é morfológica e estrutural, no sentido que a estrutura desses cristais 2D não possui periodicidade em z. Porém, nas fases em camadas das HOIP RP e DJ, a intercalação de camadas de octaedros faz surgir efeitos de confinamento que não são estruturais nem morfológicos: a diferença entre a permissividade elétrica do meio com octaedros e do meio com moléculas orgânicas espaçadoras faz surgir o confinamento dielétrico, tornando essas perovskitas poços quânticos naturais (do inglês, *quantum wells*, QW) [251, 255].

Em um QW o confinamento quântico ocorre em uma dimensão. Isso significa que os elétrons estão confinados quanticamente em 1D nas perovskitas pseudo-2D que se comportam como QW. Tradicionalmente QW são obtidos por confinamento morfológico, quando a dimensão espacial envolvida é muito pequena. De maneira geral, para que esses efeitos de confinamento quântico surjam, a dimensão Δx deve ser menor ou comparável ao comprimento de de Broglie para o movimento térmico a uma temperatura T [14]:

$$\Delta x \lesssim \sqrt{\frac{\hbar^2}{mk_bT}}, \tag{4.1}$$

onde m é a massa efetiva do elétron, k_b é a constante de Boltzmann e $\hbar$ é a constante de Dirac (constante de Planck reduzida).

O confinamento dielétrico em HOIP também faz surgir efeitos de confinamento de excítons [246, 256]. A energia de ligação do excíton em 2D E_b^{2D} está relacionada com a energia de ligação do excíton em 3D E_b^{3D} através da razão entre a permissividade elétrica ϵ_w do poço potencial e da permissividade elétrica ϵ_b da barreira de potencial através da Eq. 4.2 [256]. Uma vez $\epsilon_w > \epsilon_b$, o confinamento dielétrico deve amplificar a energia de ligação do excíton em pelo menos quatro vezes:

$$E_b^{2D} = 4E_b^{3D}\left(\frac{\epsilon_w}{\epsilon_b}\right)^2. \tag{4.2}$$

A Eq. 4.2 fornece uma estimativa de amplificação de excítons em semicondutores em camadas, mas tem limitações para HOIP por dificuldades de se medir ϵ_b das moléculas espaçadoras e ϵ_w da camada de octaedros separadamente [256].

As perovskitas híbridas metal-orgânicas, especialmente as tridimensionais (3D), como a $MAPbI_3$ (metilamônio, chumbo e iodo), apresentam notável eficiência optoeletrônica, porém sofrem de instabilidade significativa frente à umidade, calor e luz, limitando sua aplicação comercial em dispositivos fotovoltaicos [257]. Em contraste, as perovskitas híbridas em camadas (pseudo-2D), que possuem espaçadores orgânicos volumosos entre as camadas inorgânicas, mostram maior estabilidade ambiental devido à sua natureza hidrofóbica [258]. No entanto, essa maior estabilidade geralmente vem acompanhada de uma mobilidade de carga reduzida e menor eficiência quântica, evidenciando o compromisso entre desempenho e durabilidade [246]. Por exemplo, a formação de excítons na absorção é ruim para conversão fotovoltaica, onde é necessário realizar a separação das cargas (buracos e elétrons) geradas pela absorção de fótons incidentes - i.e., as cargas excitadas devem ser convertidas em corrente elétrica nos terminais do dispositivo. Apesar da presença de excítons e não linearidades óticas não contribuir para aplicações fotovoltaicas de conversão de energia, tais comportamentos podem ser explorados para emissão de luz em LEDs, onde uma maior presença de excítons deve aumentar a intensidade da recombinação de cargas, por exemplo.

Tradicionalmente, a estabilidade das perovskitas 3D pode ser atestada por uma heurística geométrica, o fator de tolerância de Goldschmidt t, proposto em 1926 por V. M. Goldschmidt [259]. Valores fora do intervalo $0.8 \leq t \leq 1.0$ indicam distorções na estrutura que podem levar à instabilidade. Seu cálculo é efetuado considerando-se o raio iônico r_A do cátion A, r_B do cátion B, e r_X do ânion X, conforme a Equação 4.3:

$$t = \frac{r_A + r_X}{\sqrt{2}(r_B + r_X)}. \tag{4.3}$$

Uma vez que a Eq. 4.3 considera apenas informações atômicas, o cálculo da predição da estabilidade da perovskita torna-se uma tarefa de rápida verificação. Posteriores aperfeiçoamentos dessa heurística tem sido feitos, entre os quais, usando restrições geométricas

[260] e aprendizado de máquina [261]. *Filip* e *Giustino* propõem certos limites (limites químicos, CL1 e CL2; limite do octaedro, OL; limite de estiramento, SL; limite de estiramento secundário, SSL1 e SSL2; e limite de angulação, TL1 e TL2) que determinam uma região PR no espaço definido por t e pelo fator octaédrico $\mu = \frac{r_B}{r_X}$, onde estruturas do tipo perovskita podem se formar [260]. Já *Bartel et al.* usaram um algoritmo de aprendizado de máquina que realiza regressão simbólica, o SISSO (Sure Independence Screening and Sparsifying Operator) para encontrar um novo descritor de estabilidade [261]. O fator de tolerância de Bartel τ é expresso por:

$$\tau = \frac{r_X}{r_B} - n_A \left(n_A - \frac{\frac{r_A}{r_B}}{\ln\left(\frac{r_A}{r_B}\right)} \right), \tag{4.4}$$

onde n_A é o estado de oxidação do átomo A e $\tau < 4.18$ indica formação de perovskita.

Na literatura, o método geométrico de *Filip* e *Giustino* alcançou uma acurácia de 80% para pervoskitas e perovskitas duplas, enquanto que o método de Bartel obteve 90% como limite inferior para perovskitas de haleto com cloro [260, 261].

Para as perovskitas híbridas, as moléculas de MA^+ e FA^+ são aproximadas como íons de raio 217 pm e 253 pm, respectivamente [262]. Moléculas espaçadoras grandes, contudo, fogem da definição de raio iônico aceitável, limitando a aplicação do fator de tolerância de Goldschmidt e da tolerância de Bartel. Apesar deste método ter reportado sucesso de classificação de estabilidade para $(MA)_2AgBiBr_6$ e pervoskitas com composição $A_2BB'X_6$, ele não generaliza, a princípio, para estruturas pseudo-2D com espaçadores.

Neste capítulo, demonstramos o uso da abordagem convencional para estudar perovskitas HOIP pseudo-2D e 3D com composição $(HDA)(MA)_{n-1}Pb_nBr_{3n+1}$ e $(MA)PbBr_3$, conforme apresentado na Figura 4.1. Essas perovskitas foram sintetizadas e caracterizadas experimentalmente, com simulações computacionais sendo realizadas na sequência. Através de espectros de fotoluminescência e absorção, e de simulações DFT, exploramos características óticas desses materiais, evidenciando a importância do confinamento quântico e dielétrico, assim como a manifestação de Efeito Rashba[1]. Uma série de possíveis investigações futuras são sugeridas a partir dos resultados encontrados e discutidos a seguir.

Ao aprofundar essa metodologia, queremos destacar os contrastes discutidos entre os Capítulos 2 (Abordagem Convencional) e 3 (Ciência de Dados de Materiais). Além disso, o domínio das técnicas convencionais é essencial para a ampla compreensão da problemática central (a descoberta de novos materiais e suas aplicações tecnológicas). Não obstante, as pervoskitas híbridas metal-orgânicas constituem um rico braço de pesquisa para os dispositivos optoeletrônicos como células solares e LEDs, sendo oportuna uma rápida excursão por essa família de materiais.

[1] O efeito Rashba-Dresselhaus refere-se ao fenômeno de acoplamento spin-órbita em certos materiais, onde o spin do elétron está acoplado ao seu momento devido à assimetria de inversão estrutural (efeito Rashba) e à simetria cristalina intrínseca (efeito Dresselhaus), resultando em comportamentos eletrônicos dependentes do spin.

4.1 Metodologia

4.1.1 Metodologia Experimental

Foi utilizada uma rota por precipitação em solução para produzir as perovskitas pseudo-2D na forma de pó (amostra policristalina). A síntese é representada através das reações químicas dadas pelas Equações 4.5, 4.6 e 4.7, enquanto para perovskita 3D é representada pela Eq. 4.8:

$$\mathrm{PbO_{(sol)} + HBr_{(liq)} + HDA_{(sol)} \rightarrow HDA\text{-}PbBr_{4(sol)}} \tag{4.5}$$

$$\mathrm{2\,PbO_{(sol)} + HBr_{(liq)} + MABr_{(sol)} + HDA_{(sol)} \rightarrow HDA\text{-}(MA)Pb_2Br_{7(sol)}} \tag{4.6}$$

$$\mathrm{3\,PbO_{(sol)} + HBr_{(liq)} + 2\,MABr_{(sol)} + HDA_{(sol)} \rightarrow HDA\text{-}(MA)_2Pb_3Br_{10(sol)}} \tag{4.7}$$

$$\mathrm{PbO_{(sol)} + HBr_{(liq)} + (MA)Br_{(sol)} \rightarrow (MA)PbBr_{3(sol)}} \tag{4.8}$$

O procedimento constituiu na dissolução de óxido de chumbo (PbO) em pó em solução de ácido bromídrico (HBr), sob agitação em chapa térmica com temperatura de 110 °C. O brometo de metilamônio (MABr) e/ou a hexametilenodiamina (HDA) foram adicionados a solução transparente de óxido de chumbo diluído, formando um precipitado. Este precipitado, em cada caso, constitui o material alvo sendo sintetizado. O material resultante é obtido por filtragem e colocado para secar em papel filtro coberto. Quando secas, as perovskitas sintetizadas foram guardadas em Eppendorf.

A síntese de monocristais de $(MA)PbBr_3$ ($n = \infty$) pode ser executada por uma rota diferente, como a cristalização por temperatura inversa [263]. Outros procedimento de síntese para $(HDA)PbBr_4$ ($n = 1$) são reportados em [245, 249] (policristalina) e [264] (monocristalina e filme fino).

Para a síntese de solução coloidal de $(HDA)PbBr_4$ ($n = 1$), o procedimento constitui em diluir óxido de chumbo (PbO) em pó em solução de dimetilformamida (DMF), sob agitação em temperatura ambiente (27 °C). Quando completamente homogênea, foi adicionado dibrometo de hexametilenodiamina ($HDABr_2$), deixando solubilizar até a solução ficar completamente transparente. Em seguida foram adicionados os surfactantes ácido oleico (OA) e oleilamina (OLAM), mantendo a agitação por mais dois minutos. Por fim, 40 μL da solução foram dispersos em um frasco com 4 mL de tolueno. A reação é sumarizada na Eq. 4.9:

$$\mathrm{PbO_{(sol)} + (HDA)Br_{2(sol)} + OA_{(liq)} + OLAM_{(liq)} \rightarrow (HDA)PbBr_{4(liq)}} \tag{4.9}$$

As amostras na forma de pó produzidas foram caracterizadas por difração de raios X

para a resolução da estrutura cristalina. Foi utilizado um difratômetro Stoe STADI-P da Bruker com linhas de radiação CuK-α_1 ($\lambda = 1.54056$) Å. Para o espectro de absorção das amostras em pó e em solução coloidal foi utilizado o espectroscópio UV-VIS Evolution 220 da Thermo-Fisher. Os espectros de PL das amostras em pó foram obtidos com espectrofotômetro de fotodiodo ultravioleta-visivel, modelo Cary 50 Varian, enquanto que para a amostra em solução coloidal foi utilizado o espectrômetro laser PicoQuant Fluorescence lifetime, modelo FluoTime 300 - testes na amostra em pó $n = 1$ também foram realizados em diferentes ângulos de incidência do laser, não diferendo qualitativamente dos espectros de PL obtidos pelo espectrofotômetro.

Como branco das medidas espectroscópicas foram medidos o ar com a cubeta e lâmina no caso das amostras em pó, e a cubeta com tolueno no caso da amostra em solução coloidal. O branco foi removido das medidas pela subtração para cada ponto de energia.

Os sinais de absorbância e de intensidade de PL foram processados com filtro de Wiener, janela de 4 amostras, implementado no SciPy [265]. Em seguida, cada sinal foi normalizado seguindo escalonamento de mínimo-máximo com relação ao próprio pico de valor. O objetivo da normalização da intensidade foi permitir uma comparação qualitativa entre os espectros das diferentes amostras, sem comprometer a análise quantitativa da posição de energia (i.e., a resposta com relação a energia do fóton incidente ou emitido).

Para a análise do espectro de absorção A realizamos um *Tauc plot* [266] utilizando o software Origin, versão 2018. Foi tomado o quadrado da absorbância e ajustada uma reta $y = ax + b$ para a região considerada de corte (posterior a região de *Urbach tail*[2]). Os coeficientes a e b ajustados pela regressão foram usados para calcular a intersecção com o eixo de energia E para determinar o bandgap $E_g = -\frac{b}{a}$. Para corroborar a análise, foi plotado $\frac{dA}{dE}$ em função de E, com o pico do gráfico correspondendo ao valor de E_g.

4.1.2 Metodologia Computacional

Para o estudo das pervoskitas pelo viés computacional empregamos técnicas de DFT utilizando a implementação VASP, versão 5.4.4, funcionais PBE-GGA. Realizamos a relaxação estrutural, a resolução da estrutura eletrônica e o cálculo do espectro de absorção.

A relaxação estrutural visa resolver a posição de menor energia dos íons da rede cristalina conjuntamente com a minimização da energia eletrônica. Dessa forma, nesta etapa relaxamos a estrutura cristalina original permitindo todos os graus de liberdade (posição dos íons, volume e forma da célula unitária) e considerando as forças de van der Waals (Método DFT-D3 de Grimme com função de amortecimento zero). Como critério de convergência eletrônica utilizamos tolerância de 10^{-5} eV e uma diferença de energia entre os passos de -0.01 eV para convergência ionica. A energia de corte para a expansão das ondas planas foi de 450 eV , com os pseudopotenciais padrão dos VASP, e uma malha de

[2] *Urbach tail* é a cauda de decaimento exponencial na curva de absorção [267] que ocorre depois da região de corte do espectro (região na qual a tendência do coeficiente de absorção α com relação a energia E é $\alpha \propto E^{1/2}$ [14]). Essa mudança de comportamento da absorção está associada a materiais amorfos e defeitos em estruturas cristalinas.

25 kpoints. Não consideramos estados de spin, nem o acoplamento spin-órbita (SOC).

Como as funções de onda são expandidas para uma base que depende da estrutura inicial, após a convergência do cálculo (passos iônicos e eletrônicos), a etapa de relaxação é repetida, isto é, o cálculo é refeito para a estrutura resultante. Esse processo é repetido até que apenas um passo iônico seja dado ao realizar a simulação DFT.

Para a obtenção do padrão de difração de raio-x de pó foi utilizado o módulo interno do VESTA [136] sobre a estrutura resultante da relaxação.

Com a estrutura cristalina relaxada, realizamos um cálculo de densidade de estados (DOS) agora com uma malha de 50 kpoints automática, mesma energia de corte, com correções de van der Waals, e considerando estados de spin e o acoplamento spin-órbita. Para o cálculo das bandas eletrônicas utilizamos a densidade eletrônica calculada e o VASPKIT [139] para gerar os caminhos de alta simetria sobre a estrutura relaxada. As massas efetivas foram obtidas com ajuste parabólico através do software sumo [268].

Para o cálculo do espectro de absorção utilizamos os mesmos critérios utilizados para o cálculo da DOS.

O coeficiente de absorção α foi calculado com relação a parte real (ϵ_1) e imaginária (ϵ_2) da função dielétrica para diferentes energias E do fóton incidente usando o módulo do pymatgen [269], sendo c a velocidade da luz no vácuo e h a constante de Planck:

$$\alpha = \frac{2\pi E}{hc}\sqrt{\sqrt{\epsilon_1^2 + \epsilon_2^2} - \epsilon_1}. \tag{4.10}$$

4.2 Resultados e Discussão

A perovkista 3D (MA)$PbBr_3$ tem fator de tolerância de Goldschmidt $t = 0.896$ e fator octaédrico $\mu = 0.731$, dento do intervalo de estabilidade. Esses valores foram obtidos através da Eq. 4.3 para $r_X = r_{\mathrm{Br}^{1+}} = 182$, $r_B = r_{\mathrm{Pb}^{2+}} = 133$ pm (coordenação VI devido as 6 ligações dentro do octaedro) e $r_A = r_{\mathrm{MA}^{1+}} = 217$ pm. O fator de tolerância de Bartel $\tau = 3.7$, calculado a partir da Eq. 4.4, também indica estabilidade.

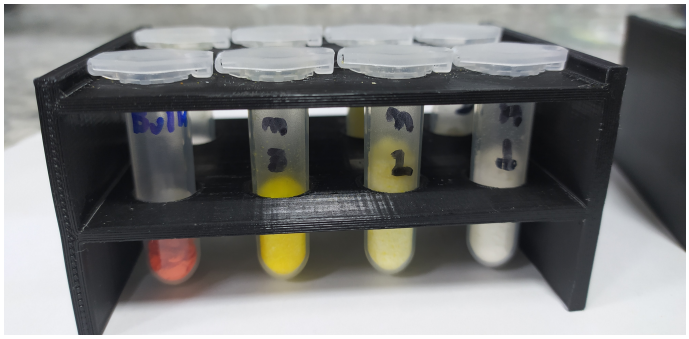
Para as perovskitas pseudo-2D (HDA)$(\mathrm{MA})_{n-1}\mathrm{Pb}_n\mathrm{Br}_{3n+1}$ temos uma mistura no sítio A, que pode ser ocupado por MA ou HDA. Contudo, se assumirmos que as extremidades das moléculas de HDA que estão ancoradas nas camadas de octaedros são grupos amina assim como as moléculas de MA, podemos extrapolar os mesmos fatores de tolerância do caso 3D para esses casos p2D.

Uma vez que estas amostras puderam ser sintetizadas, a estabilidade[3] empírica condiz com aquela predita pelas heurísticas geométricas calculadas acima.

Na Figura 4.2 são apresentados exemplos das amostras produzidas pela rota solvotermal. Essa amostras formam um pó de coloração característica: cor branca no caso da p2D n = 1, de cor amarelada no caso da p2D n = 2 e n = 3, e de cor laranja no caso da 3D.

[3]Por estabilidade nos referimos apenas a condição de *sintetizabilidade* e não a degradação com o tempo de exposição a condições ambientais como umidade

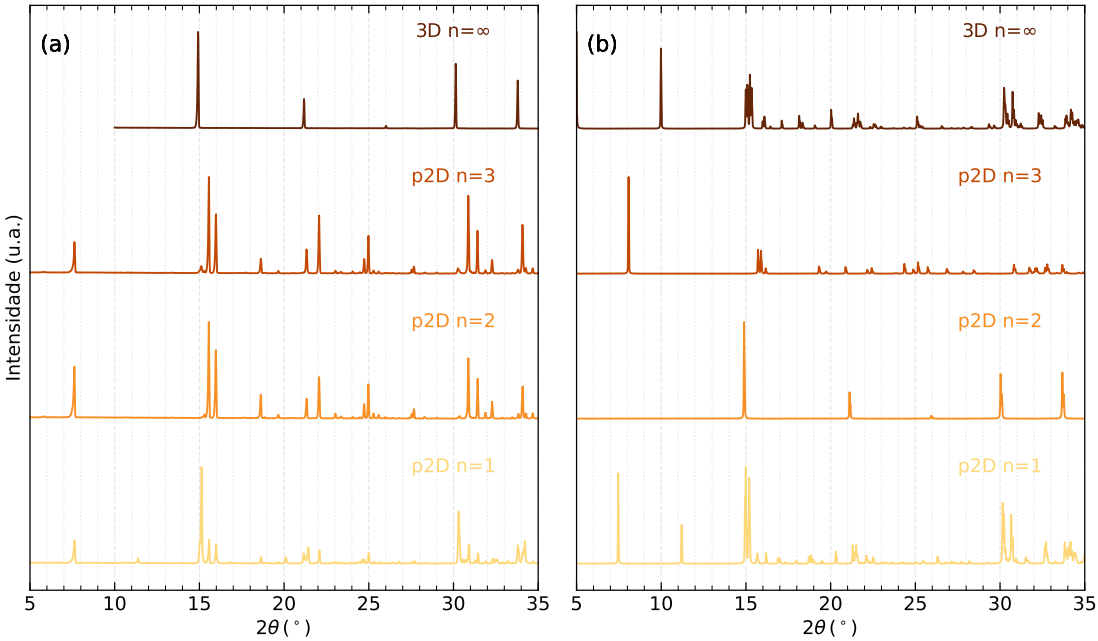
Figura 4.2: Amostras de HOIP sintetizadas na forma de pó.



Fonte: elaborado pelo autor.

Na Figura 4.3 (a) estão os espectros de DRX das amostras sintetizadas. Para comparação, em (b) apresentamos os espectros obtidos computacionalmente de estruturas CIF construídas artificialmente em uma célula cúbica e relaxadas por DFT. Observa-se nos resultados experimentais que as amostras obtidas para as HOIPs p2D são misturas de fases, provavelmente de $n = 1$ e $n = 2$ - apesar do pico teórico de $n = 3$ em $2\theta = 8°$ poder estar relacionado ao pico experimental em $2\theta = 7.5°$, os resultados de PL (Figura 4.4) não apresentam um pico característico para $n = 3$. A comparação entre o DRX experimental e computacional da p2D $n = 1$ demonstra picos nas posições corretas (por exemplo, próximos a $2\theta = 7.5°$, $2\theta = 11°$ de menor intensidade no experimental, $2\theta = 15°$ e $2\theta = 30°$).

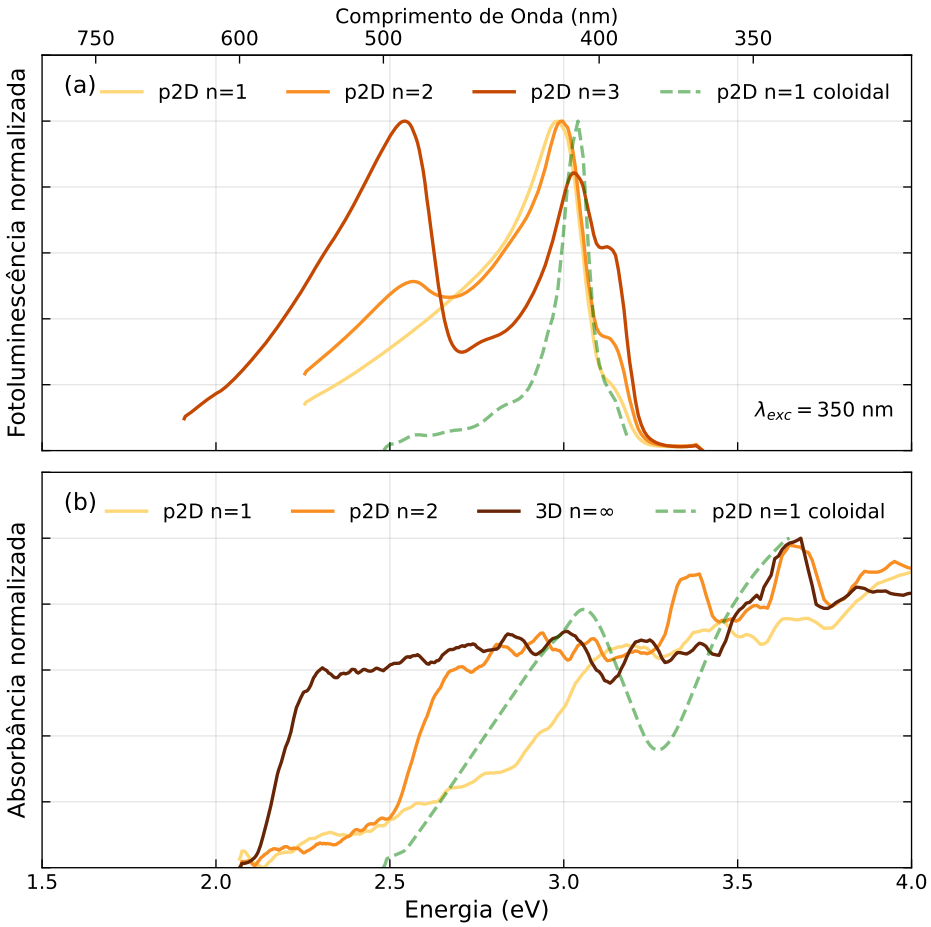
Figura 4.3: Espectro de difração de raios X das perovskitas. (a) Dados experimentais. (b) Dados computacionais da estrutura relaxada com DFT.



Fonte: elaborado pelo autor.

A Figura 4.4 contém os espectros de emissão por fotoluminescência (PL) e absorção (Abs.) de UV-VIS. Para o PL em (a), observamos um pico largo (para menores energias) em 2.98 eV para HOIP p2D $n = 1$ na forma de pó e um pico estreito em 3.04 eV para $n = 1$ em solução coloidal. Esse mesmo pico aparece para $n = 2$ em 2.99 eV e para $n = 3$ em 3.02 eV. Dada a resolução dos equipamentos e da análise de sinais, podemos considerar que este pico ocorre para todos os casos em aproximadamente 3.0 eV e é característico de $n = 1$. Já para $n = 2$ e $n = 3$ um novo pico está presente em 2.57 eV e 2.54 eV, respectivamente. Novamente, vamos considerar que se trata de um mesmo pico em 2.5 eV e que é característico das amostras $n = 2$. Esses picos nas amostras em pó $n = 1$, $n = 2$ e $n = 3$ confirmam a mistura de fase evidenciada pela análise do DRX. Também se observa no espectro de PL que há outro pico próximo a 3.14 eV (400 nm), assemelhando-se a um ombro do pico característico de $n = 1$ em 3.0 eV, que, apesar de estar presente em todas as amostras, tem intensidade relativa ao pico principal aumentada de $n = 1$ para $n = 2$ e de $n = 2$ para $n = 3$.

Figura 4.4: Espectro de emissão por fotoluminescência (a) e de absorção por UV-VIS (b) das amostras HOIPs.



Fonte: elaborado pelo autor.

Para o espectro de absorção na Fig. 4.4 (b) tanto a extrapolação da linha de corte sobre a abscissa quanto o método das derivadas retorna $E_g = 2.5$ eV para $n = 2$ e $E_g = 2.1$ eV para $n = \infty$. Na literatura, o valor reportado para (MA)$PbBr_3$ ($n = \infty$) é de $E_g = 2.0$ eV [241], muito próximo do valor observado. Para $n = 1$, o método do *Tauc plot* retorna $E_g = 2.8$ eV e o método da derivada retorna $E_g = 2.99$ eV. Como essa amostra tem a maior *Urbach tail*, consideramos que o método da derivada traz um melhor resultado.

Para a amostra $n = 1$ em solução coloidal a análise do espectro de absorção requer um pouco de atenção. De fato, há um pico de absorção em 3.054 eV (2.966 eV pelo método da derivada) que coincide aproximadamente com o pico de emissão PL em 3.04 eV na Fig. 4.4 (a). Porém, a absorção retoma um comportamento similar a $\alpha \propto E^{1/2}$ para energias mais altas. Isso corresponde ao comportamento de amostras com excítons [14]. Pelo *Tauc plot* dessa região obtemos $E_g = 3.289$ eV (3.131 eV pelo método da derivada). A diferença entre os picos é a energia de ligação do excíton, que vale aproximadamente $E_b = 0.235$ eV pelo método do *Tauc plot* e $E_b = 0.165$ eV pelo método da derivada. Na literatura, *Mousdis et al.* obtiveram $E_g = 3.346$ eV e $E_b = 0.1664$ pelo espectro de absorção de amostras $n = 1$ em temperatura ambiente [249], o que é um resultado na mesma direção do aqui apresentado - contudo, as amostras de *Mousdis et al.* estavam na forma de pó e aqui estão em solução coloidal.

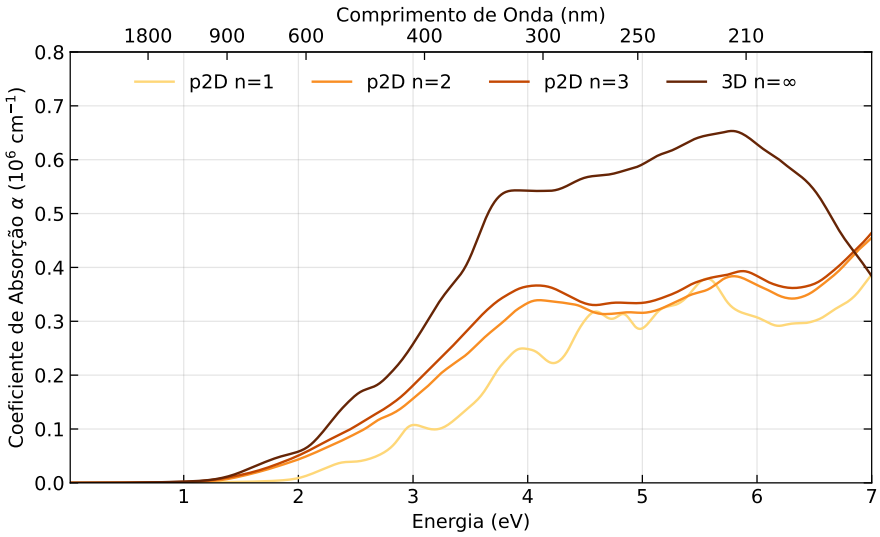
Do ponto de vista teórico, simulamos em DFT a resposta dielétrica do material, considerando sua forma *bulk* monocristalina. Tal assunção baseia-se na ideia de que este resultado deve corresponder a forma em pó policristalina obtida experimentalmente desde que o tamanho de grão destas amostras não seja da ordem nanoscópica onde podem surgir efeitos de confinamento. Para (MA)$PbBr_3$ ($n = \infty$) foi utilizada a estrutura polimórfica para o cálculo dos coeficientes ópticos e eletrônicos. Os resultados das simulações de absorção são apresentados na Figura 4.5. Observa-se para as amostras p2D $n = 1$, $n = 2$ e 3D a mesma tendência apresentada na Figura 4.4 (b), isto é, a absorção inicia-se em energia mais baixas dos fótons incidentes para o material 3D, seguido de $n = 2$ e $n = 1$. Os valores de energia, contudo, estão deslocados horizontalmente já que para o espectro de absorção pelos coeficientes α, o gap de energia é subestimado pela metodologia DFT.

Quanto a inclinação das curvas é importante lembrar que, de acordo com a Lei de Beer-Lambert, a absorbância A é uma função linear da concentração c (mol/m^3), da espessura ótica l (m) e absortividade molar ε (m^2/mol), $A = \varepsilon c l$; já o coeficiente de absorção α é dado por $\alpha = \varepsilon c$. Logo, a relação entre as curvas é dada por:

$$A = \varepsilon c l = \alpha l. \tag{4.11}$$

Ou seja, a multiplicação de α no gráfico da Figura 4.5 por uma constante $l > 1$ aumenta a inclinação das curvas. Reciprocamente, a divisão de A no gráfico da Figura 4.4 (b) por uma constante $l > 1$, diminui a inclinação das curvas. Como é esperado que l seja igual para todas as amostras, os comportamentos observados na teoria e no experimento podem ser considerados compatíveis.

Figura 4.5: Espectro de absorção das HOIPs representado pelo coef. α obtido por DFT.



Fonte: elaborado pelo autor.

Como nesses cálculos DFT estamos considerando apenas a resposta eletrônica[4] $\epsilon_\infty = \epsilon_1$, temos os tensores dielétricos para os materiais simulados como sendo:

$$\epsilon_\infty^{3D} = \begin{bmatrix} 3.5031 & 0.0261 & -0.0151 \\ 0.0261 & 3.5027 & -0.0153 \\ -0.0151 & -0.0153 & 3.4106 \end{bmatrix} \tag{4.12}$$

$$\epsilon_\infty^{n=1} = \begin{bmatrix} 3.6598 & -0.0002 & 0.0050 \\ -0.0002 & 3.7753 & -0.0934 \\ 0.0050 & -0.0934 & 3.7905 \end{bmatrix} \tag{4.13}$$

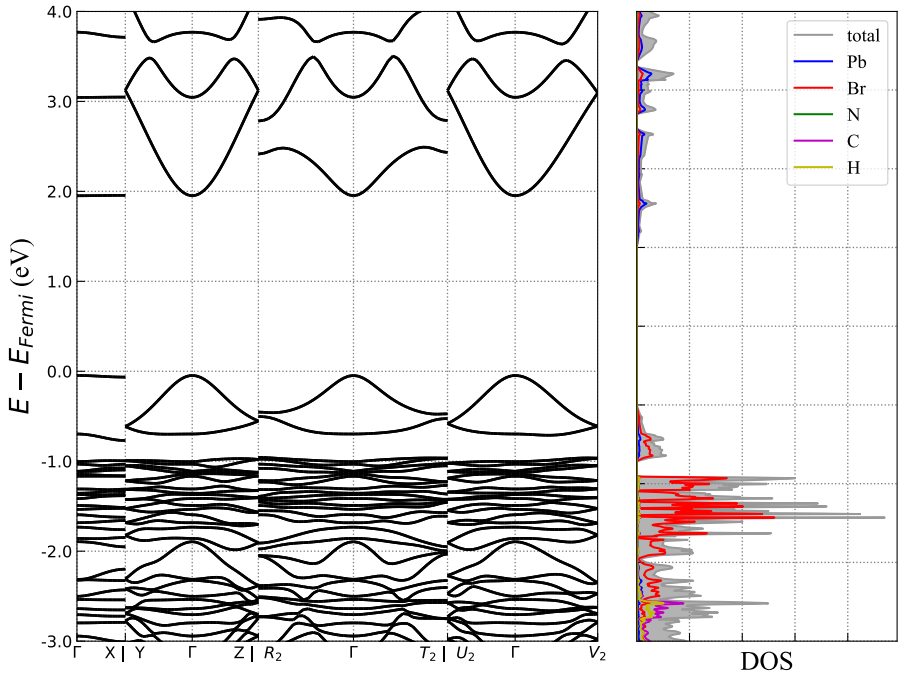
$$\epsilon_\infty^{n=2} = \begin{bmatrix} 4.1888 & 0.0315 & -0.0181 \\ 0.0315 & 4.3104 & -0.0279 \\ -0.0181 & -0.0279 & 3.6594 \end{bmatrix} \tag{4.14}$$

$$\epsilon_\infty^{n=3} = \begin{bmatrix} 4.4050 & 0.0220 & -0.0146 \\ 0.0220 & 4.4605 & -0.0201 \\ -0.0146 & -0.0201 & 3.8948 \end{bmatrix} \tag{4.15}$$

Ainda com simulações DFT, resolvemos a estrutura eletrônica para $n = 1$, $n = 2$ e $n = 3$, conforme apresentado nas Figuras 4.6, 4.7 e 4.8. Para $n = 1$ o gap de energia de $E_g = 2.00$ eV ocorre em uma transição direta no ponto gama; Para $n = 2$ e $n = 3$ os gaps são indiretos com $E_g = 1.28$ eV e $E_g = 1.14$ eV, respectivamente. Para todos os casos, os estados no topo da banda de valência são majoritariamente associados ao bromo (Br), enquanto que o estados no fundo da banda de condução são associados ao chumbo (Pb).

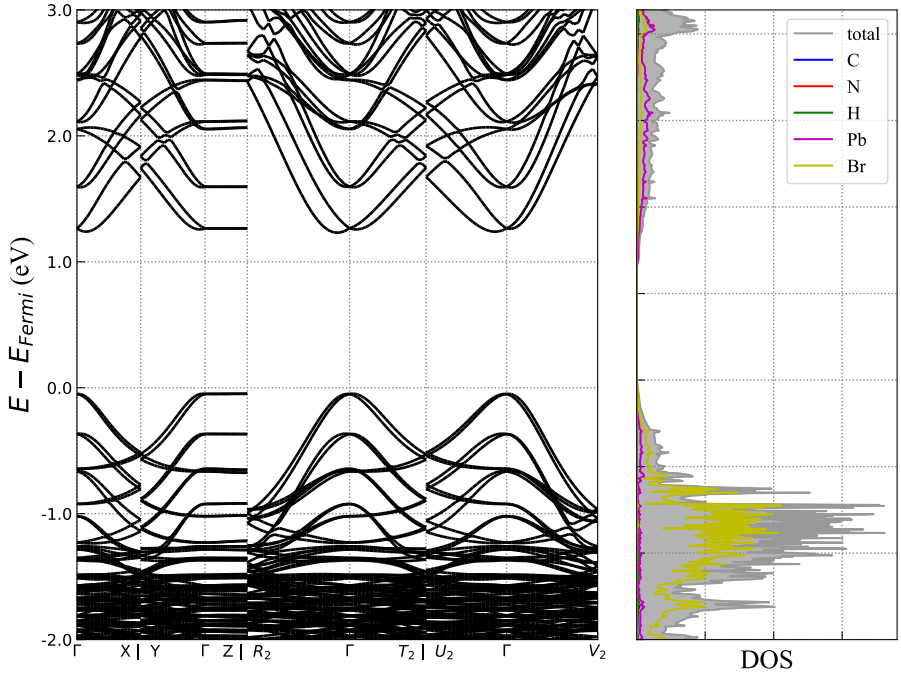
[4] O cálculo da resposta iônica para a constante dielétrica estática ϵ_e, requer DFT com perturbação.

Figura 4.6: Estrutura eletrônica e densidade de estados para HDA-$PbBr_4$, p2D $n = 1$, usando PBE-GGA com SOC. Gap de energia direto no ponto Γ, $E_g = 2.00$ eV.



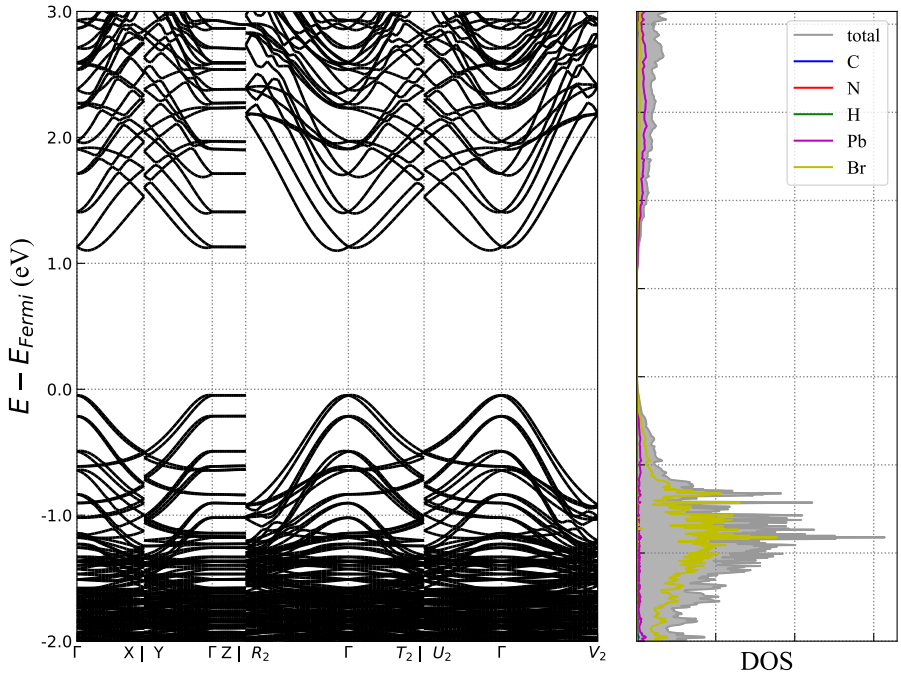
Fonte: elaborado pelo autor.

Figura 4.7: Estrutura eletrônica e densidade de estados para HDA-(MA)Pb_2Br_7, p2D $n = 2$, usando PBE-GGA com SOC. Gap de energia indireto próximo a Γ, $E_g = 1.28$ eV.



Fonte: elaborado pelo autor.

Figura 4.8: Estrutura eletrônica e densidade de estados para HDA-$(MA)_2Pb_2Br_{10}$, p2D $n = 3$, usando PBE-GGA com SOC. Gap de energia indireto próximo a Γ, $E_g = 1.14$ eV



Fonte: elaborado pelo autor.

É conhecido que o método de DFT utilizado (PBE-GGA com SOC) subestima os valores de bandgap quando comparados com os valores experimentais [270]. Além do erro sistemático esperado devido ao método PBE-GGA, devemos lembrar que o gap determinado por PL e Abs. são gaps óticos, enquanto que o gap previsto por DFT é o gap eletrônico. O gap ótico pode envolver interações de muitos corpos além da transição eletrônica entre a banda de valência e a banda de condução do gap eletrônico. Por exemplo, excítons e fônons podem estar envolvidos nos processos de emissão e absorção, podendo alterar o valor de energia observado no fóton emitido ou absorvido pelo material [271]. Por outro lado, a própria estrutura eletrônica pode sugerir a necessidade de outras partículas para a emissão e absorção de fótons. Nas Figuras 4.7 e 4.8 os gaps observados são indiretos, o que sugere que a transição eletrônica necessita da troca de momento k pela interação com fônons da rede cristalina.

Os gaps indiretos vistos nas Figuras 4.7 e 4.8 também indicam um desdobramento das bandas conhecido como desdobramento Rashba (do inglês, *Rashba splitting*). O desdobramento Rashba é a separação dos estados de spin dos elétrons em bandas de energia devido ao acoplamento spin-órbita causado pela quebra de simetria de inversão estrutural, sendo um dos componentes do fenômeno mais amplo conhecido como efeito Rashba-Dresselhaus. Em HOPIs de chumbo, o acoplamento spin-órbita leva ao Efeito Rashba-Dresselhaus, enquanto que em QW de semicondutores convencionais, o desdobramento Rashba tende a ser muito menor do que nestas perovskitas [246].

O parâmetro Rashba α_R pode ser calculado através de:

$$\alpha_R = \frac{2E_R}{k_R}, \tag{4.16}$$

onde a diferença de energia $E_R = |E(k_R) - E(0)|$ é definida entre o pico (ou vale) das bandas desdobradas e o ponto de cruzamento entre essas bandas de spin trocado, e $k_R = |k(E_R) - k(0)|$ é a diferença de momento entre as posições de pico (ou vale) e de cruzamento (conhecido em inglês como *momentum offset*).

Verificamos a quebra de simetria dentro dos octaedros de $PbBr_6$: para $n = 1$, a distância $\overline{\text{PbBr}} = 3.0\,\text{Å}$ o ângulo médio $\widehat{\text{BrPbBr}} = 180°$; para $n = 2$ átomos diferentes tiveram posições ligeiramente diferentes, de forma que foram medidos $\overline{\text{PbBr}} \approx 2.9\,\text{Å}$ e $\overline{\text{PbBr}} \approx 3.1\,\text{Å}$, com um ângulo médio de $\widehat{\text{BrPbBr}} \approx 169.7°$; para $n = 3$ observamos o mesmo comportamento de $n = 2$ para as distâncias interatômicas, com um ângulo médio de $\widehat{\text{BrPbBr}} \approx 168.5°$. Essas medidas demonstram que ocorre uma deformação do octaedro para $n = 2$ e $n = 3$, sendo essa distorção responsável por um momento de dipolo elétrico líquido e pela quebra de simetria de inversão estrutural. Como a estrutura do material deixa de ser centrossimétrica (o átomo de Pb não está mais no centro do octaedro), observamos o efeito Rashba.

A partir da análise da estrutura de bandas, calculamos o parâmetro Rashba através da Eq. 4.16 obtendo $E_R = 26.2$ meV, $k_R = 0.0422\,\text{Å}^{-1}$ e $\alpha_R = 1.24\,\text{eVÅ}$ para $n = 2$ e $E_R = 28.8$ meV, $k_R = 0.0420\,\text{Å}^{-1}$ e $\alpha_R = 1.37\,\text{eVÅ}$ para $n = 3$. Em ambos os casos, resolvemos o parâmetro Rasbha para a banda de condução. Para a banda de valência, a resolução de energia não foi suficiente para determinar a diferença entre os picos e o cruzamento das bandas desdobradas em spin.

Esses resultados teóricos sugerem que as perovskitas p2D $n = 2$ e $n = 3$ podem demonstrar efeitos spintrônicos, que abrem novas oportunidades de aplicação. Há, entretanto um ponto de discussão importante: as moléculas de MA que ocupam o sítio A dentro das camadas de octaedros em $n = 2$ e $n = 3$ estão ordenadas ao longo da estrutura, i.e., todas apontam na mesma direção. O maior estiramento da ligação Pb-Br dentro do octaedro ocorre na direção perpendicular as camadas de octaedros, provavelmente devido a penetração da molécula espaçadora HDA, que está invertida em cada lado das camadas. Logo, uma pergunta a ser respondida é se uma distribuição da orientação das moléculas de MA ao longo da camada não pode neutralizar a distorção causada pela penetração da molécula espaçadora HDA ou se a orientação observada pode ser de fato induzida naturalmente quando a estrutura está relaxada. Responder isso deve corroborar se de fato o efeito Rashba é experimentalmente observável nesses materiais, uma vez que numa condição natural, é provável que as moléculas estejam desordenadas por efeitos de temperatura e entropia - lembramos que o método DFT calcula a estrutura relaxada para $T = 0$ K e próximo da estrutura originalmente sugerida. Para obter isso de um ponto de vista teórico, o uso de Dinâmica Molecular *ab initio* pode ser um caminho, mas limitações quanto a quantidade de átomos e elétrons simulados podem surgir, uma vez que supercélulas grandes serão

exigidas para capturar toda a estrutura e as variações de orientação das moléculas.

Outro parâmetro obtido pela análise da estrutura da bandas é a massa efetiva dos portadores de carga (elétrons e buracos). Nas Figuras 4.6, 4.7 e 4.8, observa-se que as bandas têm alta dispersão indicando que a massa efetiva dos portadores é pequena. Aproximando as bandas por uma parábola, podemos calcular a massa efetiva m^* a partir do ajuste dos coeficientes da relação de dispersão $E(k)$:

$$E(k) = \frac{\hbar^2 k^2}{2m^*} + \alpha_R. \tag{4.17}$$

A princípio o vetor de onda k deve ser tomado em alguma direção, de forma que de fato há k_x, k_y e k_z. Como uma simplificação podemos assumir a massa efetiva média $< m^* >$ calculada a partir dos caminhos no espaço reciproco. Esses resultados são apresentados na Tabela 4.1, onde m_e^* é a massa efetiva dos elétrons e m_h^* dos buracos. Também consideramos que um par elétron-buraco (excíton) tem massa reduzida μ_{ex} dada pela Eq. 4.18:

$$\mu_{ex} = \frac{m_e^* m_h^*}{m_e^* + m_h^*}. \tag{4.18}$$

Tabela 4.1: Massa efetiva média dos portadores de carga obtida por DFT.

n	Massa efetiva (m_e)		Massa reduzida (m_e)
	$< m_e^* >$	$< m_h^* >$	μ_{ex}
1	0.32	0.46	0.19
2	0.33	0.28	0.15
3	0.26	0.21	0.12

Fonte: elaborado pelo autor.

Agora podemos usar o modelo de excítons livres de Wannier–Mott para estimar a energia de ligação E_b dos excítons a partir dos nossos resultados teóricos. Este modelo segue a descrição de que o par elétron-buraco forma um hidrogenoide de massa reduzida μ_{ex} e raio de órbita a_{ex} [14]:

$$E_b(n) = -\frac{\mu_{ex}}{m_e \epsilon_r^2} \frac{R_H}{n^2}, \tag{4.19}$$

onde n é o estado do excíton, ϵ_r é a constante dielétrica do material e $R_H = 13.6$ eV é a constante de Rydberg para o átomo de hidrogênio.

$$a_{ex}(n) = n^2 \frac{m_e}{\mu_{ex}} \epsilon_r a_H, \tag{4.20}$$

onde $a_H = 0.529177$ Å é o raio de Bohr.

Os tensores dielétricos calculados sugerem que a anisotropia dielétrica é baixa uma vez que os valores fora da diagonal principal são muito pequenos e os valores na diagonal

principal são próximos. Logo, podemos assumir que $\epsilon_r = \frac{1}{3}\mathrm{Tr}(\epsilon_\infty)$.

Quanto a natureza do confinamento quântico, considerando a estrutura das perovskitas estudadas, temos uma espessura da camada de octaedros (poço) $d_w \approx 6\,\text{Å}$, $d_w \approx 12\,\text{Å}$, $d_w \approx 18\,\text{Å}$ para $n = 1$, $n = 2$ e $n = 3$, respectivamente. Em todos os casos a espessura da camada orgânica dada pela molécula de HDA (barreira) é $d_b \approx 7\,\text{Å}$. A Tabela 4.2 apresenta esses valores com os raios de órbita dos excítons estimados pela Eq. 4.20:

Tabela 4.2: Tamanhos estimados para as perovskitas p2D.

n	d_w (Å)	d_b (Å)	a_{ex} (Å)
1	6	7	10.5
2	12	7	14.2
3	18	7	19.4

Fonte: elaborado pelo autor.

Podemos utilizar a Eq. 4.1 para estimar o tamanho crítico no qual ocorre o confinamento quântico. Para a situação de $T = 300$ K e usando uma massa efetiva de $m^* = 0.3m_e$, esse valor é de aproximadamente $x = 3.1$ nm. Os valores de d_w e d_b na Tabela 4.2 são menores do que esse valor crítico $x_c = 31\,\text{Å}$, o que indica a possibilidade do surgimento de efeitos de confinamento quântico observáveis a temperatura ambiente. Dado que as dimensões são compatíveis, a diferença entre o confinamento quântico dielétrico e o confinamento quântico morfológico pode ser entendido como a altura da barreira potencial (profundidade do poço) devido a diferença de permissividade elétrica: neste a diferença está entre o meio orgânico e inorgânico e naquele entre o material e o vácuo. Logo, um nanocristal de (HDA)$PbBr_4$ ($n = 1$) com raio de até 31 Å deve apresentar ambos os efeitos.

A metodologia DFT empregada captura a natureza qualitativa do confinamento dielétrico que ocorre quando o número de camadas de octaedros n diminui. Tanto o resultado experimental quanto teórico concordam que a medida que os elétrons se tornam mais confinados na camada de octaedros, devido a menor espessura transversal da camada, o bandgap observado aumenta; quando o número de octaedros aumenta transversalmente, a espessura da camada aumenta e o confinamento dielétrico diminui, de forma que o valor de bandgap diminui aproximando-se do valor do material 3D (onde o número de octaedros pode ser considerado $n = \infty$). Entretanto nossos cálculos DFT não capturam efeitos de confinamento quântico, já que não simulam um cristal finito com interface com o vácuo.

Do ponto de vista experimental, entendemos que a natureza morfológica de cada caso não estava completamente resolvida. Na verdade, os nanocristais em solução coloidal ou em pó devem seguir uma distribuição de tamanho, cuja média deve corresponder aos efeitos observados. Logo precisaríamos de uma medida experimental para determinar qual o tamanho médio dos cristais em forma de pó e em solução para de fato caracterizar se as amostras podem apresentar confinamento quântico devido a morfologia. Poderia ser empregada técnica de potencial Zeta com espalhamento dinâmico de luz para determinar

o raio hidrodinâmico dos cristais em solução coloidal e microscopia eletrônica de varredura (ou transmissão) para as amostras em pó.

Na Tabela 4.3 sumarizamos os valores de bandgap e energia de ligação do excíton obtidos pelos métodos experimentais e por DFT. Assumimos que o excíton de Wannier está no estado fundamental ($n = 1$).

Tabela 4.3: Bandgap e energia de ligação do excíton determinados por métodos experimentais (gap ótico) e simulações computacionais (gap eletrônico).

Material					E_g (eV)		E_b (meV)	
Composição	Morf.	Dim.	n	PL	Abs.	DFT	Abs.	DFT †
(HDA)$PbBr_4$	coloidal	p2D	1	3.0	3.2 – 3.3	-	165 – 235	-
(HDA)$PbBr_4$	pó	p2D	1	3.0	3.0	2.00	-	183
(HDA)(MA)Pb_2Br_7	pó	p2D	2	2.5	2.5	1.28 *	-	125
(HDA)$(MA)_2Pb_3Br_{10}$	pó	p2D	3	-	-	1.14 *	-	87
(MA)$PbBr_3$	pó	3D	∞	-	2.1	1.17 ‡	-	-

*gap indireto. † modelo de Wannier. ‡cálculo do espectro de absorção.

Fonte: elaborado pelo autor.

A princípio, os excítons calculados poderiam ser observados, uma vez que a energia de ligação é mais alta que a energia térmica esperada, $k_bT \approx 0.025$ eV. Contudo, os resultados experimentais de absorção não constataram excítons livres para as amostras p2D em pó. Lembramos que *Mousdis et al.* [249] observaram excítons livres para $n = 1$ em pó com $E_b = 166.4$ meV próximo do valor estimado pelo nosso modelo teórico ($E_b = 183$ meV) e dentro do valor observado para $n = 1$ em solução ($E_b = 165 - 235$ meV). Isso levanta a questão de por que amostras de mesma composição química e morfologia (pó) apresentaram resultados diferentes de absorção com pico de excíton (*Mousdis et al.*) e sem excíton (Figura 4.4 b), enquanto que de mesma composição química, mas morfologia diferente (coloidal), apresentaram resultados compatíveis?

Primeiramente não podemos descartar que as medidas de absorção das amostras em pó sofrem de qualidade devido a perdas por espalhamento. Como o equipamento de UV-VIS utiliza parte da luz transmitida, as perdas por reflexão difusa podem diminuir a qualidade do sinal medido e esconder picos de excíton que sejam menos intensos. Para confirmar que esse não é um problema de resolução experimental dos picos, é necessário refazer os experimentos com um equipamento que tenha esfera integradora. Também é desejável realizar fotoluminescência resolvida no tempo (TRPL) para determinação dos tempos de vida radiativos (Eq. 2.62) e eficiência quântica (Eq. 2.63), assim como espectro PLE para corroborar o espectro UV-VIS.

Os espectros de absorção das amostras p2D $n = 1$ e $n = 2$ em pó demonstram *Urbach tails* consideráveis. Estas estão relacionadas a desordem térmica e estrutural [272]. Essa

região pode esconder picos de excíton que tenham uma intensidade menor. Uma modelagem mais precisa dessa região do espectro pode fornecer a energia de ativação dos defeitos, que poderiam ser corroboradas por medidas de condutividade elétrica dependente da temperatura. Não só a determinação da energia de ativação, mas também uma estimativa precisa da densidade de defeitos em cada tipo de amostra deve melhorar o processo de caraterização e ajudar na investigação da diferença observada.

Além disso, a literatura sugere que efeitos de confinamento espacial (morfológico) aumentam a energia de ligação dos excítons, o que justificaria a sua presença em um tipo de amostra e não no outro: para (MA)$PbBr_3$ são observados excítons com E_b de 40 a 70 meV para amostras *bulk* e 320 a 388 meV para pontos quânticos [272]. O mesmo poderia ocorrer para $n = 1$ em pó e em solução coloidal.

Devido a natureza da interação eletrostática que mantêm o par elétron-buraco, campos elétricos aplicados ao material podem ionizar os excítons [14, 273, 274]. Em outras palavras, um campo elétrico uniforme tende a separar os pares elétron-buraco que estão ligados via força de Coulomb. Distorções dos octaedros, como sugerem os resultados computacionais de $n = 1$ e $n = 2$, podem gerar momentos de dipolo elétrico líquido devido a assimetria no comprimento das ligações químicas. A presença desse dipolo na amostra faz surgir um campo elétrico local. Apesar de ser esperado que a orientação dos dipolos sigam uma distribuição com a temperatura, o campo local pode ser suficientemente forte para ionizar os excítons. Um estudo computacional mais detalhado das distorções dos octaedros e cálculo do momento de dipolo liquido, com posterior utilização em modelos de campo local, pode sugerir um limiar teórico de ionização do excíton (somente excítons com energia de ligação maior seriam observados). Pelos resultados aqui apresentados é possível estimar que o campo de ionização (o campo elétrico necessário para quebrar o excíton) é $|\vec{E}_c| \approx 3.5 \times 10^8$ V/m para $n = 1$, $|\vec{E}_c| \approx 1.8 \times 10^8$ V/m para $n = 2$, e $|\vec{E}_c| \approx 0.9 \times 10^8$ V/m para $n = 3$. Esses valores são o módulo do campo elétrico calculado por [14]:

$$|\vec{E}_c| = \frac{2E_b}{ea_{ex}}. \tag{4.21}$$

Os resultados da Eq. 4.21 para as energias de ligação de excíton calculadas anteriormente estabelecem que, para uma espessura de 1 μm, seria necessária uma tensão elétrica de 349 V, 176 V e 90 V, respectivamente, para a ionização desses excítons. Logo, para os valores de E_b calculados seguindo o modelo de Wannier para semicondutores 3D, a ionização do excíton pode ser pouco provável pois o campo local deve ser muito intenso e as distorção observadas não parecem corresponder a efeitos de mesma ordem. Já o efeito de duplo confinamento faria com que a energia de ligação do excíton em 2D fosse pelo menos quatro vezes maior (para (HDA)$PbBr_4$ isso seria $183 \times 4 = 732$ meV) o que não parece ser o caso.

Por outro lado, acoplamentos elétron-fônon ou buraco-fônon poderiam aprisionar o excítons livres em excítons autoaprisionados (do inglês, *self-trapped excitons*, STE). De fato os picos de PL na Fig. 1.2 (a) são assimétricos e com grande dispersão, o que pode ser

uma assinatura para esse tipo de fenômeno. Excítons autoaprisionados teriam energia de ligação maior, mas não estariam envolvidos em processos de absorção: a absorção gera os excítons livres que em seguida são aprisionados via uma distorção da rede. Medidas de PL em função da temperatura poderiam indicar o desvio do pico de PL (desvio Stokes) com relação a região de corte do espectro de absorção - para a temperatura ambiente usada aqui, esse desvio não aparece de maneira significativa. Do ponto de vista teórico, modelos de muitos corpos como método GW e correções de Bethe–Salpeter poderiam ajudar a simular a presença dos estados excitados (excítons). Contudo, é de nosso entendimento que a quantidade de átomos necessários para a simulação destas amostras tornam tais correções proibitivas computacionalmente.

4.3 Conclusão

Neste capítulo, realizamos experimentos com uma família de perovskitas pseudo-2D do tipo Dion-Jacbson, com composição $(HDA)(MA)_{n-1}Pb_nBr_{3n+1}$, e 3D, com composição $(MA)PbBr_3$. Essas perovskitas foram sintetizadas em duas formas morfológicas, dispersão coloidal de nanocristais e pó policristalino, e caraterizadas por DRX, PL e UV-VIS. Em posse dos resultados, realizamos simulações computacionais DFT para entender o espectro de absorção e a estrutura de bandas eletrônica.

A Tabela 4.3 condensa os principais parâmetros óticos observados e calculados. Do ponto de vista teórico, as composições com duas camadas de octaedros ($n = 2$) e três camadas de octaedros ($n = 3$) apresentaram Efeitos Rashba de $\alpha_R = 1.24$ eVÅ e $\alpha_R = 1.37$ eVÅ, respectivamente. Tanto resultados teóricos quanto experimentais sugerem a presença de excítons livres (excítons de Wannier-Mott) nas amostras, de acordo com outros resultados observados na literatura [249]. Contudo, limitações de resolução dos espectros de absorção e do controle de tamanhos dos nanocristais de cada a amostra, impedem uma análise conclusiva dos motivos de presença ou não dos excítons em cada amostra. Sugestões de novas iterações são feitas tanto para as questões experimentais quanto teóricas.

O principal objetivo deste capítulo foi demonstrar como a abordagem convencional recorre a simulações computacionais para corroborar e explicar resultados experimentais. Mostramos como os dados obtidos experimentalmente podem ser confrontados *a posteriori* com resultados de DFT, que, por sua vez, ajudam a esclarecer determinados mecanismos e a sugerir novos experimentos. Este processo iterativo é crescente, acumulando informações cada vez mais detalhadas sobre o material estudado e as características dos seus fenômenos intrínsecos.

Do ponto de vista da Ciência e Engenharia de Materiais, incia-se com o material conhecido, sua caraterização para depois seu estudo computacional. O problema dessa sequência ocorre quando queremos um material totalmente novo que tenha certos atributos. Por exemplo, a pergunta inicial a ser respondida deveria ser: qual material apresenta coeficiente Rashba de $\alpha_R = 1.24$ eVÅ e bandgap de $E_g = 2.5$ eV? De um ponto de vista da Ciência de Dados de Materiais a resposta, $(HDA)(MA)Pb_2Br_7$, deveria ser dada sem a necessidade

da síntese prévia do material, pois um sistema racional de predição consegue estabelecer tal resposta. Obviamente, este material sugerido seria apenas uma das respostas possíveis, inclusive pois só foram impostas duas restrições (coeficiente Rashba e bandgap). Podemos fazer uma analogia com a construção de um prédio: não devemos construir o prédio para saber se ele atende nossos requisitos; nosso projeto do prédio deve ser acurado o suficiente para que de fato o prédio possa ser construído atendendo os requisitos.

Mesmo com as limitações dos métodos computacionais, a prerrogativa da Ciência de Dados de Materiais é que é possível utilizar metodologias racionais para orientar *a priori* o estudo de um determinado material - especialmente quando o objetivo é o design inverso do material.

Apesar disso, é importante destacar que a exploração de técnicas puramente computacionais sem o devido conhecimento dos detalhes empíricos acarretaria em uma metodologia de predição ruim ou profundamente limitada. Se voltarmos a analogia com a construção do prédio, um bom projetista deve entender de obras para não propor projetos inexequíveis.

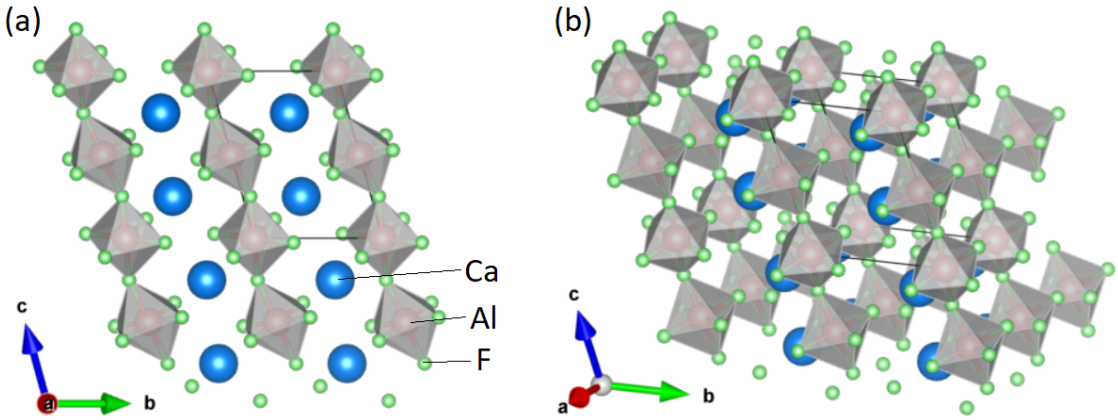
No próximo capítulo exploramos um conjunto de técnicas para o design inverso, baseado em preceitos da Informática de Materiais e dos cálculos *high throughput*.

5 Estudo de Caso II: Jakobssonita e Perovskitas pseudo-1D

A Jakobssonita, $CaAlF_5$, é um mineral encontrado em crostas de fumarolas vulcânicas na Islândia [275], recebendo esse nome em homenagem a Sveinn Jakobsson, o primeiro a reportá-la em 2008 [276]. Apesar da descoberta de sua formação natural ser mais recente, a síntese química do pentafluoreto de cálcio e alumínio já era conhecida pelo menos desde 1923 [277, 278]. De fato, fluoritos cristalinos e vidros de fluoreto, como $CaAlF_5$ e similares, foram exaustivamente pesquisados de 1971 a 1985 [279, 280, 281, 282, 283]. Em 1991, o refinamento da α-$CaAlF_5$ foi relatado por A. Hemon e G. Courbion, demonstrando que essa estrutura é isotípica com a estrutura do $CaCrF_5$ [284]. Mais recentemente, Rehmer *et al.* relataram uma rota de síntese da Jakobssonita pelo método sol-gel, em vez da síntese térmica em estado sólido conhecida até então [278].

Na Figura 5.1 é apresentada a estrutura α da Jakobssonita. Os octaedros de flúor formam cadeias conectadas pelos vértices ao longo do eixo c. Essas estruturas de octaedros $(MF_6)_n$ em zigue-zague são uma classe de compostos de fluoreto com a fórmula $A^{II}M^{III}F_5$, semelhante a um corte em uma estrutura de perovskita [283, 285].

Figura 5.1: Representação do cristal de α-$CaAlF_5$ em duas orientações (a) e (b).



Fonte: elaborado pelo autor.

O $CaAlF_5$ apresenta uma transição de fase cristalina reversível $\alpha \leftrightarrow \beta$ em torno de 740 °C [277]: β é uma fase de alta temperatura, isotípica com $CaFeF_5$, apresentando simetria monoclínica $P2_1/c$ [286] e cadeias de octaedros similares; α também é monoclínica, mas com simetria $C2/c$ [284]. P. Hagenmuller relatou cerca de 32 estruturas isotípicas de fluoretos sólidos AMF_5, sendo 7 com a jakobssonita (α-$CaAlF_5$, α-$CaGaF_5$, $CdCoF_5$,

$CaCoF_5$, $CaVF_5$, $CaCrFe_5$, $CaTiF_5$) e 6 com β-$CaAlF_5$ (β-$CaAlF_5$, β-$CaGaF_5$, β-$MnGaF_5$, $CdCrF_5$, $CdGaF_5$,$CaFeF_5$) [283].

Do ponto de vista computacional, L. Curtiss e A. Heinricher utilizaram simulações *ab initio* para estudar os materiais $BeAlF_5$ e $MgAlF_5$ com interesse na formação de complexos de halogeneto metálico MLX_5 em vapores de alta temperatura durante reações químicas, concluindo que a estrutura com grupo pontual C_{2v} é a mais termodinamicamente estáveis para esses compostos [287]. As moléculas de $CaAlF_5$ em fase gasosa foram estudadas por L. Joubert *et al.* usando análise topológica da função de localização eletrônica (ELF) em uma abordagem de DFT [288]. Diferentemente de Curtiss e Heinricher, Joubert *et al.* estavam interessados nos vapores que aparecem em processos metalúrgicos de alumina envolvendo fusões mistas de cloreto e fluoreto, e seus resultados sugerem a simetria C_{3v} em vez da C_{2v} como a estrutura molecular mais estável. Por outro lado, H. Inoue *et al.* simularam espectros Raman de α-$CaAlF_5$, β-$CaAlF_5$ e outros vidros e cristais fluoroaluminatos usando Dinâmica Molecular, mostrando a ocorrência de três bandas como resultado de vibrações de estiramento de octaedros AlF_6, em concordância com resultados experimentais [289].

Neste capítulo usamos técnicas de *screening* para selecionar a jakobssonita em bases de dados de materiais computacionais. Em seguida, utilizando substituição química, criamos 100 novos compostos com a mesma estrutura da α-$CaAlF_5$, sendo que 80 destes compostos mantiveram-se na mesma fase estrutural após a relaxação por DFT. Para cada composto dessa nova família, calculamos as fases competidoras, cujas estruturas foram obtidas originalmente de bases de dados, e posteriormente expandidas via substituição química. Ao total foram realizados aproximadamente 3 mil simulações computacionais seguindo a abordagem de HT-DFT, incluindo análise de estabilidade termodinâmica e estrutura eletrônica dessa nova família de compostos.

Com essa metodologia, exploramos a intersecção entre a Informática de Materiais e a Simulação Computacional de Materiais, avançando no sentido das técnicas específicas da Ciência de Dados de Materiais. Nosso objetivo foi demonstrar como criar compostos em experimentos *in silico* através de uma busca exaustiva e orientada por heurística química.

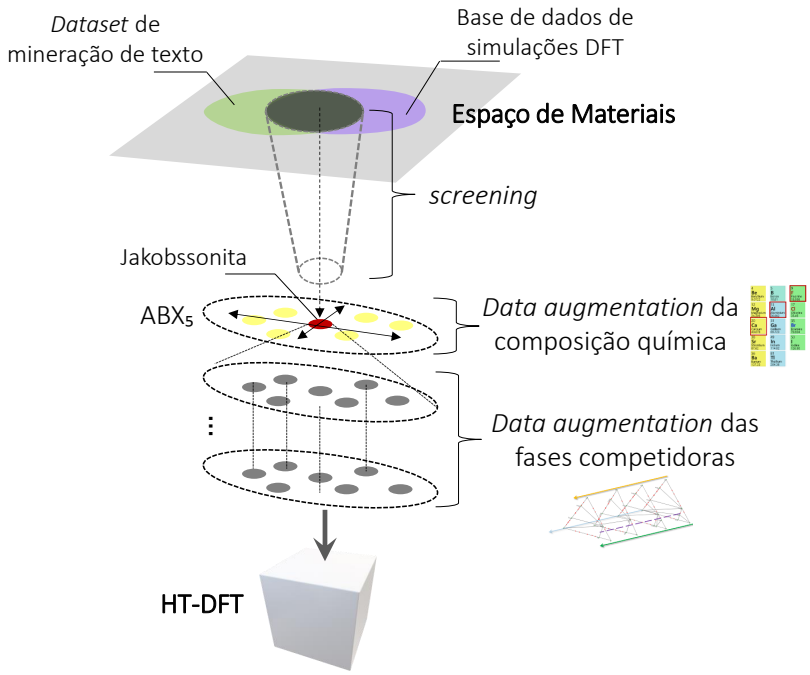
5.1 Metodologia

5.1.1 Espaço de Busca

Uma visão geral da nossa metodologia é esquematizada na Figura 5.2. Começamos nossa busca no espaço de todos os materiais cristalinos inorgânicos. Utilizamos o conhecimento prévio de estruturas desse tipo contidas na base de dados computacional Materials Project, cruzando informações com o *dataset* de síntese de materiais construído por mineração de texto reportado em [175]. Com esse conjunto de dados realizamos um *screening* para selecionar candidatos a aplicação em LEDs emissores no UV. Com o candidato selecionado, a jakobssonita, aplicamos uma etapa de aumento de dados, primeiro crescendo o espaço de busca através da geração de novas estruturas por substituição química e, em se-

guida, repetindo o processo sobre as fases competidoras. Por fim, propriedades de interesse foram calculadas usando métodos quânticos e métodos heurísticos.

Figura 5.2: Diagrama dos processos de busca de novos materiais.



Fonte: elaborado pelo autor.

O problema da sintetizabilidade foi dividido na determinação da estabilidade termodinâmica e da rota de síntese. Como heurística para resolver a questão da rota, utilizamos conhecimento de domínio ao limitar o espaço de busca original durante o *screening* da estrutura protótipo, onde escolhemos materiais com ocorrência experimental e rota de síntese reportada. Isso não é garantia de que as novas estruturas geradas, mesmo que estáveis, serão sintetizáveis pela mesma rota de síntese, porém é uma informação que pode ser utilizada como ponto de partida.

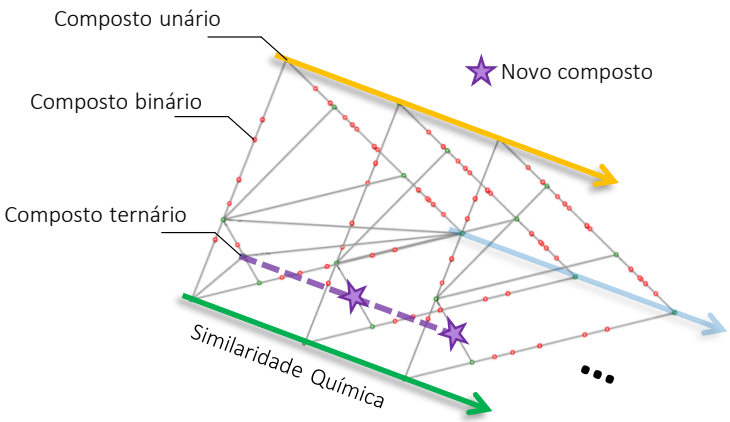
Considerando uma abordagem do tipo gere-e-teste o problema de predição de estruturas cristalinas (CSP) pode ser resolvido por dois módulos: um que gera estruturas e outro que testa sua estabilidade. Em nossa metodologia, o módulo gerador não é completo, no sentido que não gera todas as estruturas possíveis - isso ocorre pois não variamos as posições atômicas continuamente para cada composição química. Além disso, o espaço estrutural ficou limitado a estrutura protótipo selecionada pelo *screening* e a busca no espaço composicional foi limitada em $n = 14$ elementos (metais alcalinoterrosos, grupo do boro e halogêneos). Com essas restrições de arranjo e considerando que o material protótipo é ternário, temos $\Omega = 5 \times 5 \times 4 = 100$. Assim o aumentado de dados baseado na combinação química resultou na geração de novas 99 estruturas possíveis.

Entretanto, este espaço não é suficiente para atestar a estabilidade. Na verdade, a física

do problema traz uma inter-relação entre os estados possíveis: além dos materiais estarem em um mínimo de energia com relação a posição dos seus átomos e elétrons constituintes, para um material ser estável com relação ao ambiente químico, ele precisa estar em um mínimo de energia com relação a outros compostos. Ou seja, não basta testar a estabilidade do material alvo criado pelo gerador, é preciso testar a estabilidade dos outros materiais que podem ser encontrados no mesmo ambiente químico da solução proposta.

Para resolver isso, tivemos que expandir nosso espaço de busca através das fases competidoras: realizamos um *screening* nas bases de dados Materials Project e OQMD por todos os materiais que pertencessem aos 100 sistemas alvo e selecionamos estequiometrias únicas daqueles tivessem a menor energia de formação - ou seja, limitamos o espaço estrutural. Em seguida, realizamos um *data augmentation* no espaço composicional projetando cada estrutura de uma fase competidora existente em um sistema para outro sistema alvo, conforme esquematizado na Figura 5.3. Para cada triângulo de fase (sistema alvo), os pontos representam materiais (estrutura e composição): pontos sobre as arestas correspondem a materiais binários; pontos sobre os vértices, materiais unários; e pontos internos, a materiais ternários. Excetuando-se materiais unários (todos previamente existentes), quando um material não existia em um triângulo mas existia em outro, projetamos a estrutura existente sobre o novo triângulo substituindo os respectivos átomos. Em caso de existir mais de um material protótipo (um ponto em mais de um triângulo mas não sobre todos), realizamos a seleção da estrutura protótipo calculando um ranking de similaridade entre os elementos químicos originais e os da nova estrutura baseado na escala de Pettifor modificada [134]. Essa escala determina a probabilidade de um elemento químico substituir outro em estruturas binárias. Por exemplo, o flúor tem 23% de probabilidade de ser substituído por hidrogênio, 14.8% por cloro, 9.4% por bromo, 9.8% por iodo, 3.4% por oxigênio e 2.1% por enxofre. Ou seja, usando esse ranqueamento, se uma estrutura AB−F fosse a projeção de uma estrutura existente entre AB−Cl e AB−Br, seria escolhida AB−Cl como protótipo.

Figura 5.3: Diagrama do processo de aumento de dados de fases competidoras.



Fonte: elaborado pelo autor.

Como módulo testador utilizamos métodos quânticos de HT-DFT, considerando que tais métodos são baseados no conhecimento físico do problema e são capazes não só de testar a solução CSP como gerar outras propriedades de interesse. É importante destacar que esse método não é somente um testador da estabilidade, já que ele executa um processo de relaxação da estrutura, isto é, uma busca local[1] no espaço contínuo das posições atômicas, do qual uma das propriedades resultantes é exatamente a energia da estrutura convergida. Esse módulo testador também nos fornece informação sobre a estrutura de bandas, que constitui uma das propriedades de interesse de aplicação.

A seguir detalhamos as considerações físicas que utilizamos para configurar as simulações computacionais realizadas por este módulo testador para explorar o espaço de busca que construímos nesta etapa.

5.1.2 Simulações Computacionais HT-DFT

Neste trabalho nós usamos a versão 5.4.4 do software de simulação *Viena Ab initio Simulation Package* (VASP) que implementa a metodologia DFT [290]. Foram utilizados funcionais PBE-GGA e pseudopotenciais PAW para reduzir o esforço computacional visando o ganho no volume de cálculos realizados.

Configuramos uma energia de corte de 800 eV após testes preliminares de convergência. Não consideramos efeitos magnéticos, portanto não dividimos a densidade eletrônica por spin. Para a etapa de relaxação das posições dos íons permitimos mudanças de posição, formato da célula e volume da célula, com critério de convergência de tolerância de 10^-5 eV e diferença de energia entre passos de -0.01 eV. Não incluímos correções de van der Waals uma vez que não estamos trabalhando com materiais com moléculas ou em camadas. Amostramos os pontos do espaço reciproco de maneira automática com $k = 25$ pontos para a relaxação, $k = 40$ para cálculo da densidade de estados e $k = 50$ pontos para o cálculos do espectro de absorção. Assim como no caso da perovskita (Capítulo 4), todo o processo de relaxação das fases jakobssonita e das fases competidoras foi realizado de maneira iterativa, até que apenas um passo iônico fosse realizado. Porém neste caso, não incorporamos efeitos de SOC pois não tratamos de átomos pesados como de chumbo.

Para o cálculo das estruturas de bandas dos materiais da família jakobssonita, modificamos a amostragem do espaço recíproco para seguir um caminho específico de alta simetria, mantendo a mesma energia de corte dos passos anteriores. Utilizamos os caminhos reportados por [131], recuperados de maneira automática do site AFLOW utilizando um *script* de *web scraping*. Para o cálculo do espectro de absorção, utilizamos a mesma metodologia descrita no Capítulo 4, sem utilizar correções de SOC e de van der Waals, e mudando a energia de corte para 800 eV.

[1] Com local nos referimos ao fato que o cálculo parte da estrutura inicial fornecida, de maneira que outras inicializações poderiam render outros caminhos pelo espaço de posições para cada átomo, podendo resultar em outros mínimos.

5.2 Resultados e Discussão

5.2.1 Screening e Aumento de Dados

Inicialmente foi buscado na base de dados Materials Project materiais que não tivessem elementos radioativos, que possuíssem bandgap entre 3.1 eV e 12.4 eV, com entrada reportada na base de dados experimental ICSD e com estrutura de banda calculada. Como resultado foi obtida uma lista com 7094 materiais.

Na sequência, foi feita a intersecção entre os materiais desta lista e os materiais reportados por Kononova et al. [175] minerados na literatura via técnicas de NLP e cuja rota de síntese era descrita. Como resultado foi obtida uma lista reduzida de 404 materiais.

Destes 404, foram selecionados materiais únicos que tivessem bandgap direto, rendendo uma lista com 48 compostos que reproduzimos na Tabela 5.1:

Tabela 5.1: Lista de materiais obtidos pelo screening.

MP ID	**Fórmula**	**Grupo Espacial**	**# Rotas**	**Bandgap (eV)**
mp-555495	K3Y(BO3)2	58	1	3.18
mp-27377	K4Zr5O12	164	1	3.19
mp-4408	Sm2Zr2O7	227	3	3.1
mp-6943	Ca4GdB3O10	8	1	3.32
mp-13133	K2TiO3	63	4	3.41
mp-559574	K3Sm(BO3)2	62	1	3.42
mp-6687	Ba2HoSbO6	225	1	3.42
mp-5977	Nd2Zr2O7	227	4	3.33
mp-19330	CaMoO4	88	6	3.5
mp-3952	BaY2O4	62	1	3.25
mp-561536	NaMgBO3	15	1	3.59
mp-6037	Ba2SmSbO6	225	1	3.68
mp-6660	BaNaBO3	12	1	3.65
mp-10444	Ca2Al2O5	46	1	3.57
mp-3788	SrAl4O7	15	1	3.75
mp-560420	K3Bi2(PO4)3	15	2	3.78
mp-3901	NdNbO4	15	3	3.66
mp-4571	CaZrO3	62	12	3.83
mp-4974	La2Zr2O7	227	8	3.73
mp-18834	SrMoO4	88	9	3.84
mp-4709	PrNbO4	15	1	3.75
mp-4867	CaAl4O7	15	2	4
mp-17275	SrY2O4	62	3	3.7

Continua na próxima página

Tabela 5.1 – Lista de materiais obtidos pelo screening.

MP ID	Fórmula	Grupo Espacial	# Rotas	Band gap (eV)
mp-5586	NaErO2	15	1	4.05
mp-19276	BaMoO4	88	5	4.09
mp-559348	BaZr(BO3)2	167	3	4.12
mp-5487	Sr3SiO5	130	5	3.76
mp-14345	BaZnP2O7	2	2	4.34
mp-19426	CaWO4	88	12	4.35
mp-540868	Ca3Si(ClO2)2	14	2	4.38
mp-6499	BaLiBO3	14	3	4.39
mp-555087	Ba5Si(Cl3O2)2	15	1	4.55
mp-17612	Ba2SiO4	62	6	4.63
mp-19163	SrWO4	88	7	4.68
mp-556703	Ba3Mg(SiO4)2	147	2	4.69
mp-558669	SrAl2B2O7	167	2	4.69
mp-555983	Er2SiO5	15	1	4.76
mp-3520	Y2SiO5	15	5	4.77
mp-5652	Y2Si2O7	12	7	4.81
mp-4753	Al2SiO5	58	6	4.84
mp-14355	Cs2SrP2O7	15	1	4.85
mp-18803	Na2WO4	227	2	4.75
mp-561237	LiCaBO3	61	4	5.04
mp-17975	KSrPO4	62	3	5.09
mp-556532	CaAlB3O7	67	1	5.38
mp-9625	LiMgPO4	62	3	5.43
mp-10250	BaLiF3	221	3	6.57
mp-8836	CaAlF5	15	1	7.15

Fonte: elaborado pelo autor com dados de [146, 175].

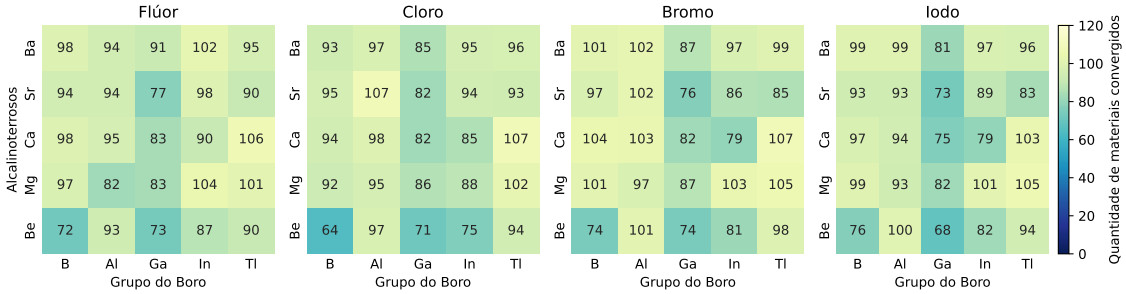
Destes resultados decidimos trabalhar com um material que não apresentasse oxigênio na composição química, restando dois candidatos finais: $BaLiF_3$ e $CaAlF_5$. Apesar do bandgap de ambos estar acima do valor desejado para a produção de LEDs UVC, nossa hipótese inicial era de que, ao trocar o flúor por outro haleto, poderíamos controlar o bandgap ajustando-o para a região de interesse. Por fim, decidimos trabalhar com o $CaAlF_5$, material cuja fase já era conhecida na literatura como Jakobssonita.

A característica isotópica desses sistemas de fluoretos é interessante para realizar substituições atômicas e obter novos compostos. Usando a jakobssonita (α-$CaAlF_5$) como protótipo, realizamos substituições nos átomos de flúor por outros halogênios (Cl, Br, I), nos átomos de alumínio por outros elementos da família do boro (B, Ga, In, Tl) e nos átomos de cálcio por outros metais alcalinos terrosos (Be, Mg, Sr, Ba). Ao fim chegamos

em um total de $A \times B \times C = 5 \times 5 \times 4 = 100$ de novas estruturas. Denominados cada uma dessas estruturas de compostos alvo, e o respectivo sistema, de sistema alvo.

Para cada composto alvo (jakobssonita SG = 15) estudado, simulamos as fases competidoras para obter todas as energias de formação e compará-las, a fim de determinar a estabilidade termodinâmica desses compostos. Para cada um dos 100 sistemas estudados havia um espaço químico possível de 120 estruturas - obtidas pelo aumento de dados de projeção sobre as diferentes combinações químicas (Figura 5.3). Além do espaço químico, cada composto alvo também possuía fases competidoras quanto a estrutura, por exemplo, a β-jakobssonita. Na Figura 5.4 são apresentadas as quantidades de estruturas convergidas por HT-DFT para cada um dos sistemas:

Figura 5.4: Contagem de estruturas convergidas (todas as fases) para cada sistema alvo.

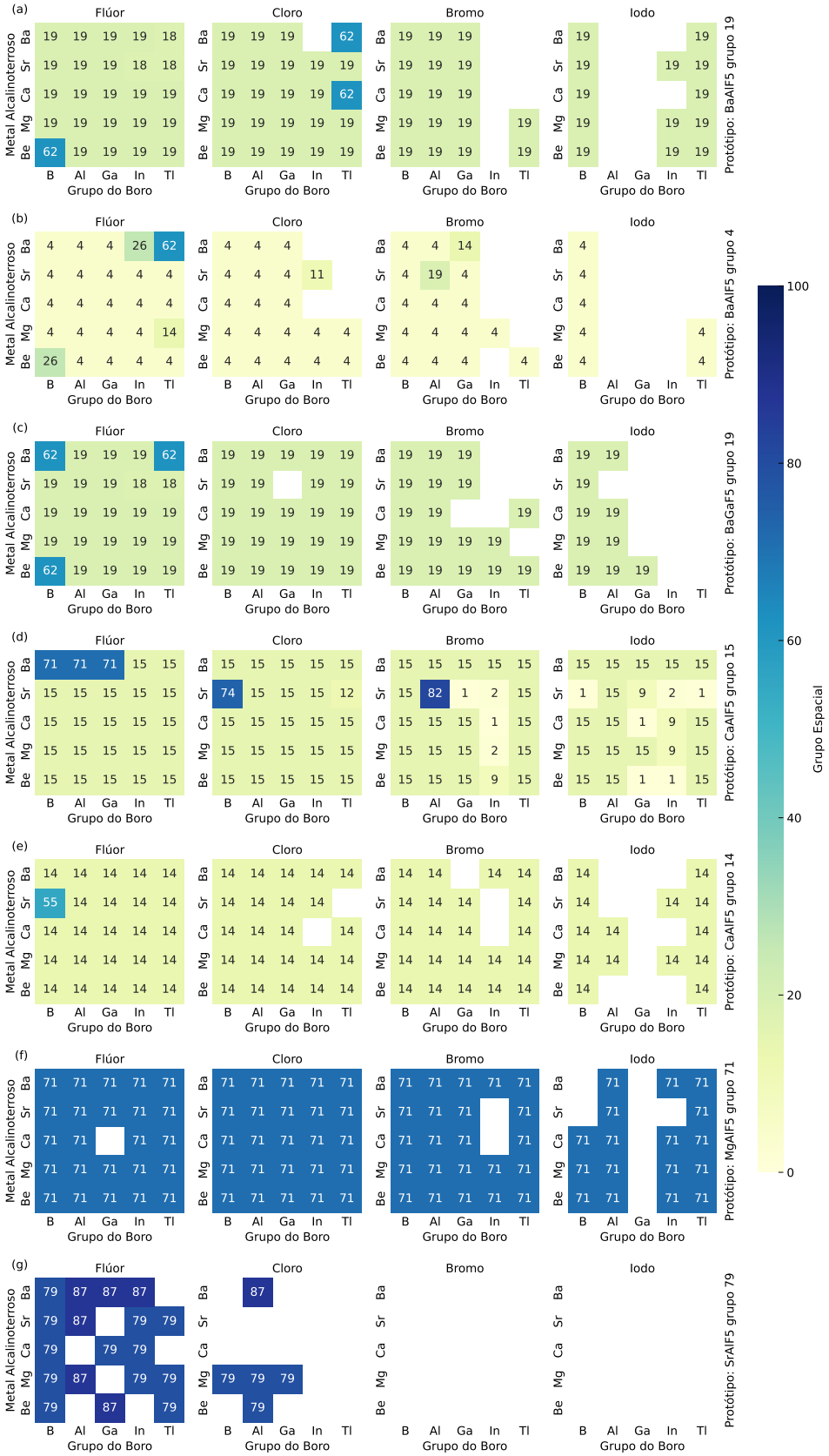


Fonte: elaborado pelo autor.

Em nenhum sistema foi possível simular os 120 compostos diferentes originais, já que algumas estruturas não convergiram. Mesmo com um tratamento manual desses casos problemáticos, por exemplo, aumentado-se o tempo de máquina para até 60 horas, não foi possível convergir todos os casos, o que pode indicar que a estrutura protótipo não era adequada para os átomos substituídos. A Figura 5.4 também mostra que compostos com alumínio ou magnésio foram os que mais convergiram e com berílio ou gálio os que menos convergiram. No total foram simuladas 2933 estruturas diferentes. É importante notar esse o total não foi de $100 \times 120 = 12000$ uma vez que parte das estruturas em um sistema também pertence a outro sistema, por exemplo AlF_3 pertence ao diagrama do sistema Be-Al-F, mas também aos sistemas Mg-Al-F, Ca-Al-F, Sr-Al-F e Ba-Al-F. Logo, isso é válidos para unários e binários, de forma que apenas os compostos ternários são unicamente pertencentes a um sistema alvo específico.

Neste total estão incluídas outras 6 famílias de compostos com a mesma composição ABC_5, totalizando 7 grupos iniciais (700 materiais), de forma a contemplar cerca variabilidade estrutural. Essas estruturas protótipo foram selecionadas no Materials Project assim como a $\alpha-CaAlF_5$ (grupo espacial número SG = 15). Foram elas a $\beta-CaAlF_5$ (SG = 14), $BaAlF_5$ (SG = 19), $BaGaF_5$ (SG = 19), $BaAlF_5$ (SG = 4), $MgAlF_5$ (SG = 71) e $SrAlF_5$ (SG = 79). Na Figura 5.5 são apresentados os grupos espacias das estrutura relaxadas quando os cálculos convergiram (521) - caso contrário, o respectivo espaço está em branco.

Figura 5.5: Grupos espaciais das estruturas alvo relaxadas por HT-DFT para cada protótipo de origem. Espaços em branco representam estruturas não convergidas.



Fonte: elaborado pelo autor.

Na Tabela 5.2 sumarizamos o total de compostos simulados agrupados de acordo com a classe de condutividade e estabilidade termodinâmica. Estes resultados se referem aos cálculos de relaxação estrutural usando PBE-GGA.

Tabela 5.2: Contagem de compostos simulados por HT-DFT (PBE-GGA) separados por tipo de condutividade e por critério de estabilidade termodinâmica.

Tipo de Condutividade	Critério de Estabilidade			Total
	$H_f > 0$	$H_f \leq 0$		
		$E_h > 0$	$E_h = 0$	
Metais $E_g = 0$	366	1199	83	1648
Semicondutores $E_g > 0$	47	1062	176	1285

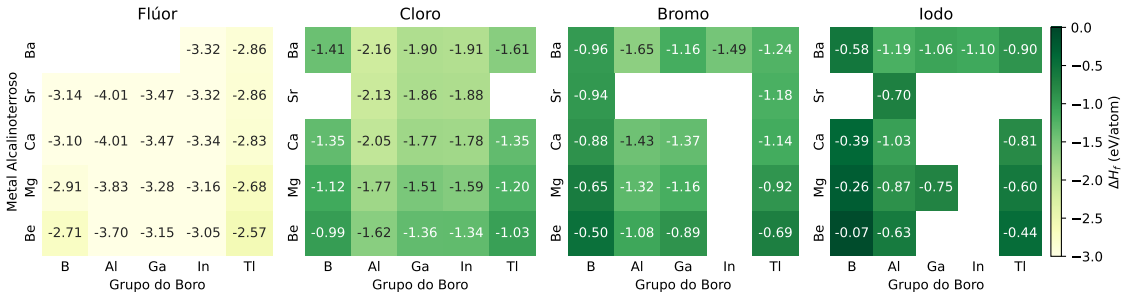
Fonte: elaborado pelo autor.

Denominamos de resultados alvo (materiais alvo) as estruturas $\alpha-CaAlF_5$ com grupo espacial 15, que totalizam ceca de 80 materiais. Para essas, realizamos uma análise um pouco mais aprofundada. Todavia, todos os outros materiais simulados constituem resultados válidos, que denominaremos de resultados secundários (materiais secundários). A seguir discutimos em mais detalhes esses dois tipos de resultados quanto a estabilidade termodinâmica e o bandgap.

5.2.2 Estabilidade Termodinâmica

Na Figura 5.6 apresentamos a entalpia de formação (Eq. 2.57) para cada um dos compostos alvo. Apesar de todos terem energia negativa, indicando sua propensão à formação, observa-se uma tendência de crescimento da energia do flúor para o iodo (mais estável para menos estável).

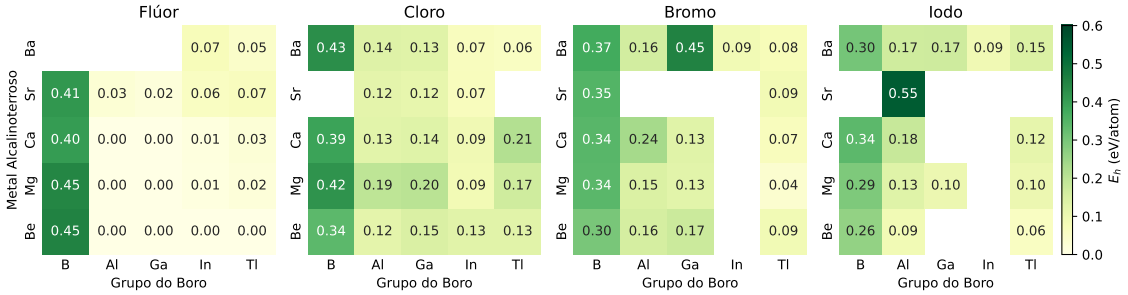
Figura 5.6: Entalpia de formação ΔH_f para cada composto alvo obtida das simulações DFT PBE-GGA da relaxação.



Fonte: elaborado pelo autor.

Quanto a estabilidade com relação as fases competidoras, na Figura 5.7 são apresentadas as distâncias de cada composto alvo para o *convex hull*. A colunas de boro mostram que esse elemento tende a formar fases menos estáveis (longe do *hull*). O flúor possui a mair quantidade de materiais no *hull* ou muito próximo dele.

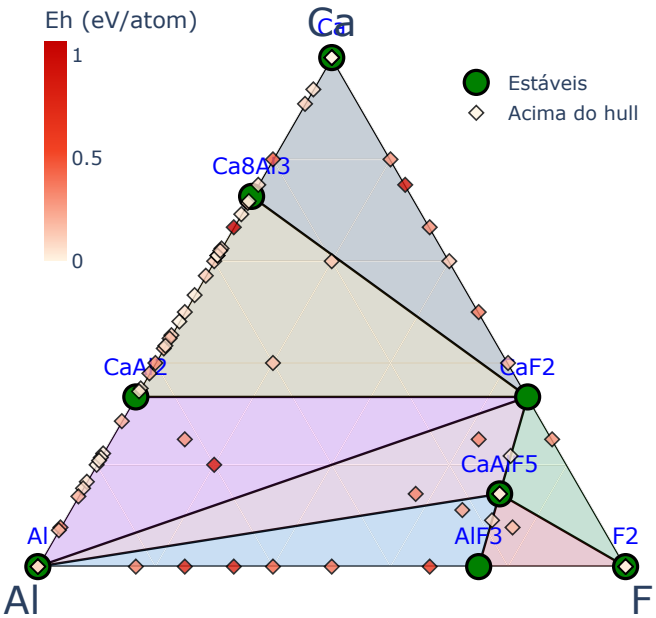
Figura 5.7: Distância de energia para o *convex hull* E_h para cada composto alvo obtida das simulações DFT PBE-GGA da relaxação.



Fonte: elaborado pelo autor.

Para que esses valores de distância do *convex hull* fossem obtidos, foram construídos os diagramas de fase para cada um dos 100 sistemas alvo - apesar de 20 deles terem sido omitidos nos gráficos das Figuras 5.6 e 5.7 já que a estrutura alvo convergiu para um grupo espacial diferente do protótipo original. Na Figura 5.8 apresentamos o diagrama de fase para o sistema Ca-Al-F. Os demais sistemas estão no Apêndice A no final deste trabalho.

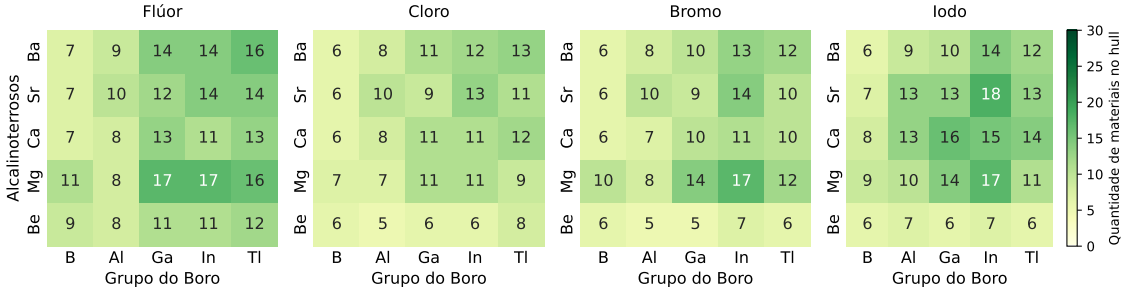
Figura 5.8: Diagrama de fase para o sistema Ca-Al-F.



Fonte: elaborado pelo autor.

O total de materiais estáveis, $E_h = 0$, para cada um dos 100 sistemas é apresentado na Figura 5.9. Comparada com o total de estruturas simuladas, a quantidade de materiais estáveis foi entre 5% e 20% considerando cada sistema isolado. No total 259 estruturas no *convex hull* para 2933 simulações equivale a 8.8% do total do espaço explorado. A tendência química observada é que o berílio e o boro normalmente são aqueles com menor número de estruturas no *convex hull* - provavelmente pois seus tamanhos relativamente pequenos com relação as outros átomos não favorece a formação de estruturas. Já entre os halogênios, o cloro se mostrou com a menor quantidade de materiais com $E_h = 0$.

Figura 5.9: Contagem de estruturas no *convex hull* para cada sistema alvo.



Fonte: elaborado pelo autor.

Devemos, entretanto, perceber que diversas estruturas estão muito próximas do *convex hull* o que as torna metaestáveis, com grandes chances de também serem encontradas na natureza ou sintetizadas em laboratório. Além disso nosso cálculos são efetuados para uma temperatura virtual nula ($T = 0$). Para uma maior compreensão dessas características, outras simulações poderiam ser feitas, principalmente as de Dinâmica Molecular, refletindo os efeitos da temperatura e da pressão sobre essa estruturas. Também efeitos dinâmicos de vibração da rede poderiam tornar o material instável ou estável, sendo necessário neste caso, cálculos de estrutura de fônons.

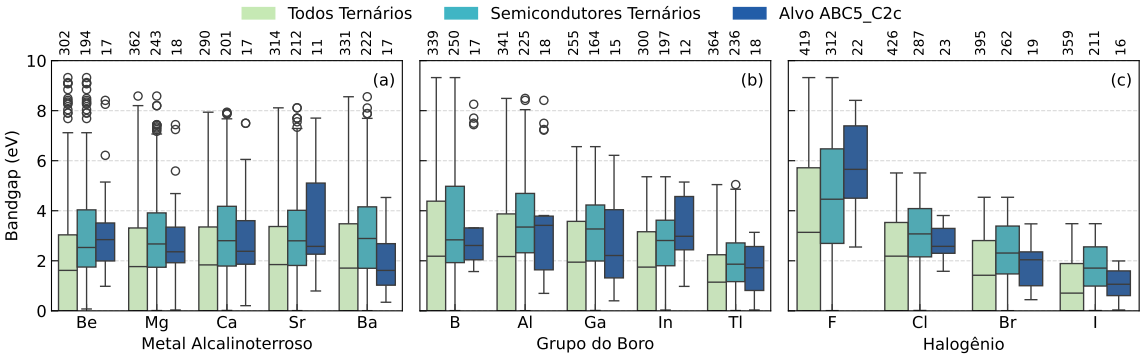
O objetivo de atestar a estabilidade termodinâmica de um material alvo (os compostos da família da jakobssonita que aqui criamos) reside apenas em um dos critérios de *sintetizabilidade* que foram aqui endereçados. Podemos relaxar esse critérios para um valor específico acima do *convex hull* com o qual ainda acreditamos que o material possa ser sintetizado. Isso é discutido na Seção 5.2.4. Por outro lado, essa busca em profundidade sobre cada sistema gerou uma série de novos compostos que podem ou não atender nosso critérios originais de aplicabilidade e *sintetizabilidade*. Esse materiais secundários são coprodutos da nossa metodologia *high throughput*, sendo uma contribuição importante para investigações futuras mais detalhadas. De fato, tais resultados compõem uma potencial nova base de dados que explora em maior profundidade os materiais formados pelos elementos dos metais alcalinoterrosos, do grupo do boro e dos halogênios - constituindo, portanto, uma demonstração de como se construir essa bases no âmbito da Informática de Materiais.

A seguir discutimos as propriedades eletrônicas apenas dos materiais alvo.

5.2.3 Propriedades Eletrônicas

Para observar a tendência química do bandgap, agrupamos os compostos pelos grupos químicos, Figura 5.10. Observa-se que, estatisticamente, os metais alcalinoterrosos não exercem controle sobre o bandgap; elementos do grupo do boro tem uma ligeira tendência de diminuição do máximo de gap; os halogênios controlam significativamente o bandgap, que é gradativamente menor na média, conforme descemos nesta coluna da tabela periódica - este é o mesmo comportamento observado em perovskitas de haletos tradicionais; o grupo de materiais alvo segue a mesma tendência de todos os semicondutores ternários com estes elementos; há presença de metais para todas as combinações desses elementos químicos.

Figura 5.10: Distribuição do bandgap E_g de todos os ternários obtida por HT-DFT da relaxação estrutural, agrupada por elemento químico. Os números indicam quantidade de compostos em cada *boxplot*.



Fonte: elaborado pelo autor.

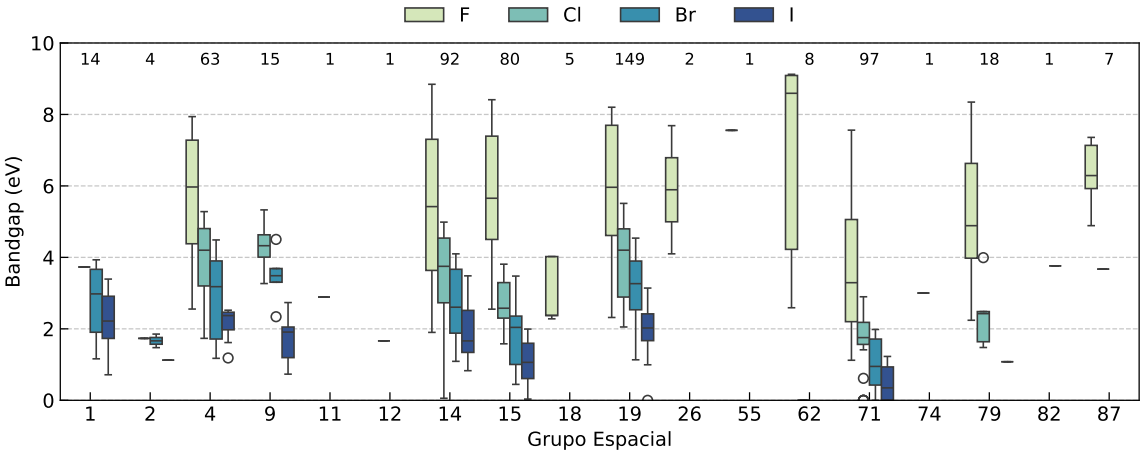
Conclui-se dessa análise que não podemos afirmar que um agrupamento químico diferente (com relação a coluna da tabela periódica) faz o material ser metal ou semicondutor, porém conseguimos atestar a tendência de controlar o gap que cada grupo químico tem - no caso, os halogênios são os principais responsáveis pela diminuição do bandgap nos ternários com essas combinações químicas.

Também verificamos a influência da simetria da estrutura para o valor de bandgap. Focando apenas nas estruturas ABC_5, agrupamos por grupo espacial e por halogênio (para isolar o fator mais importante verificado no passo anterior). Esses resultados são apresentados na Figura 5.11. Os grupos espacias com estatística mais confiáveis são aqueles que possuem maior número de materiais, sendo eles SG = 4 (63), SG = 14 (92), SG = 15 (80), SG = 19 (149) e SG = 71 (97). Dessa análise não foi possível encontrar uma tendência nítida de controle do gap pelo grupo espacial.

Já na Figura 5.12 analisamos a influência do grupo químico na diretividade do gap (se direto ou indireto) somente para os 80 materiais alvo, com resultados do cálculo DFT da estrutura de bandas. Observa-se que o berílio (a), o boro e o tálio (b), e o bromo (c) são os elementos mais presentes quando o bandgap é indireto - i.e., a frequência de materiais com bandgpap indireto é mais de 2 vezes maior que a frequência daqueles com gap direto

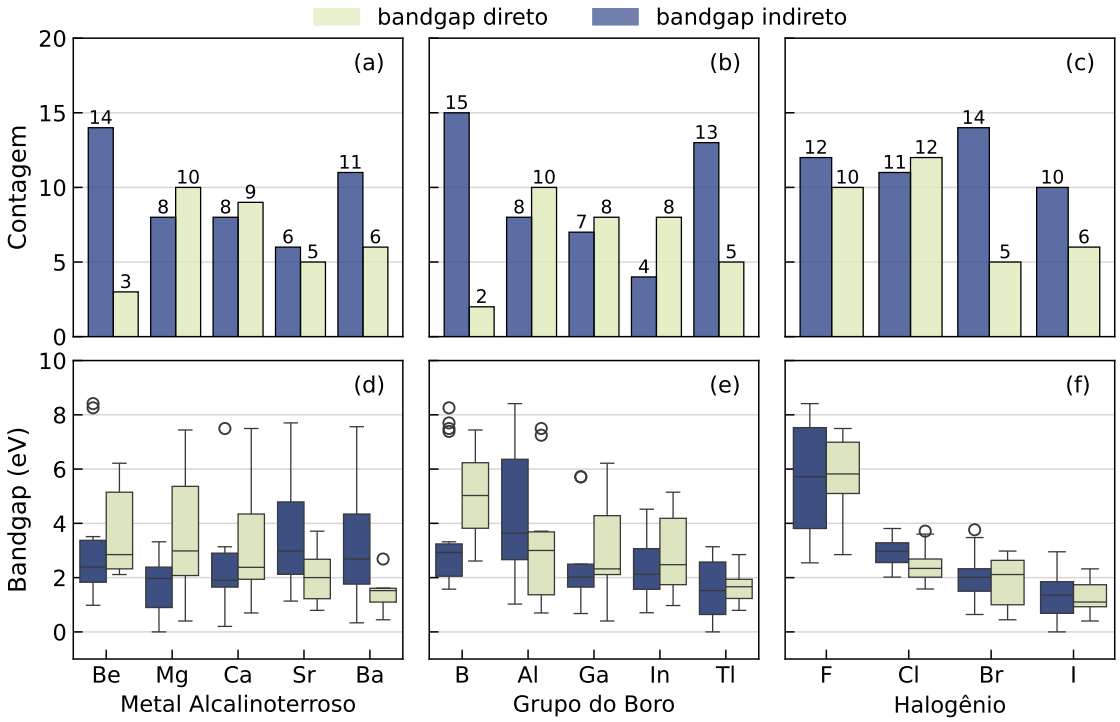
para o agrupamento com esses elementos químicos.

Figura 5.11: Distribuição do bandgap E_g dos ternários ABC_5 obtida por HT-DFT da relaxação estrutural, agrupada por grupo espacial e halogênio. Os números indicam quantidade de materiais em cada grupo espacial.



Fonte: elaborado pelo autor.

Figura 5.12: Distribuição da diretividade e energia de bandgap E_g para os materiais alvo obtida das simulações DFT PBE-GGA da estrutura de bandas.

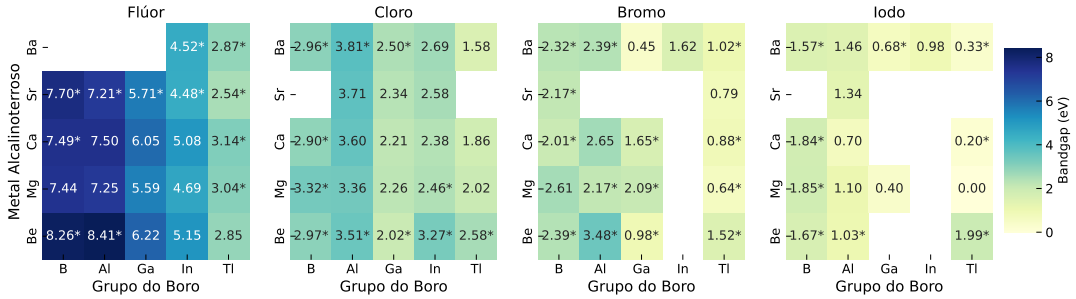


Fonte: elaborado pelo autor.

Na Figura 5.13 são apresentadas matrizes com os valores do gap de energia de cada

um dos 80 compostos alvo, incluindo informação sobre a diretividade. A mesma tendência química evidenciada nos gráficos de *boxplot* pode ser observada através da coloração ao longo das matrizes. Assim como na análise da estabilidade termodinâmica, espaços vazios correspondem a estruturas que não convergiram para o grupo espacial de interesse.

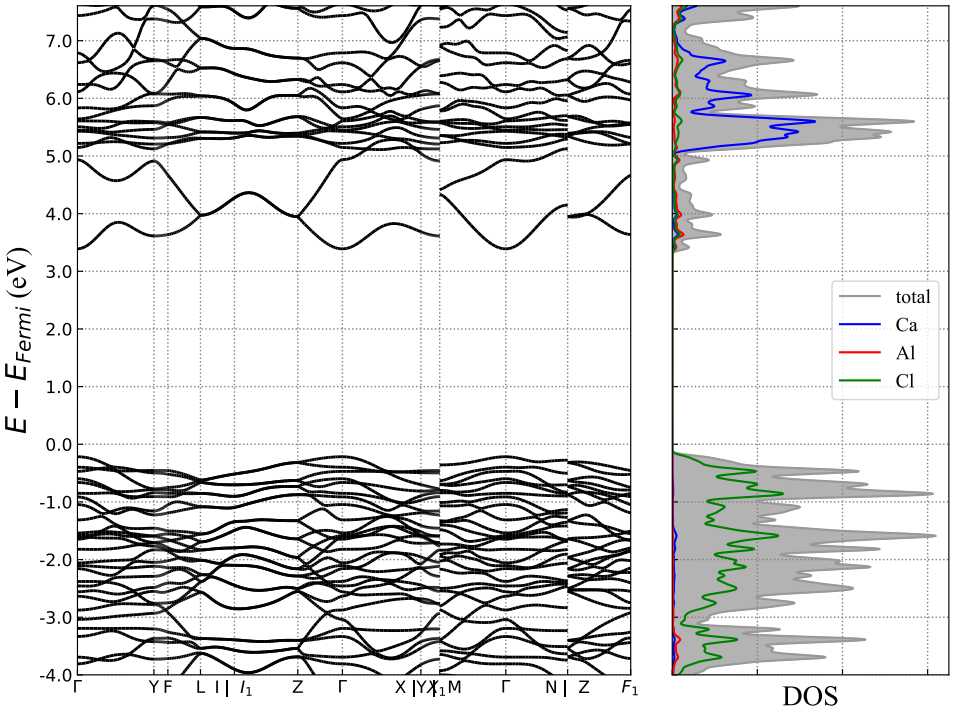
Figura 5.13: Distribuição do bandgap E_g para cada composto alvo obtido das simulações DFT PBE-GGA da estrutura de bandas. A marcação com * indica gap indireto.



Fonte: elaborado pelo autor.

Para a obtenção dos dados presentes na Figura 5.13 foram realizados cálculos de estrutura de bandas e de densidade de estados para os 80 materiais alvo. Esses resultados são mostrados na Figura 5.14 para o $CaAlCl_5$. Os demais gráficos estão no Apêndice B.

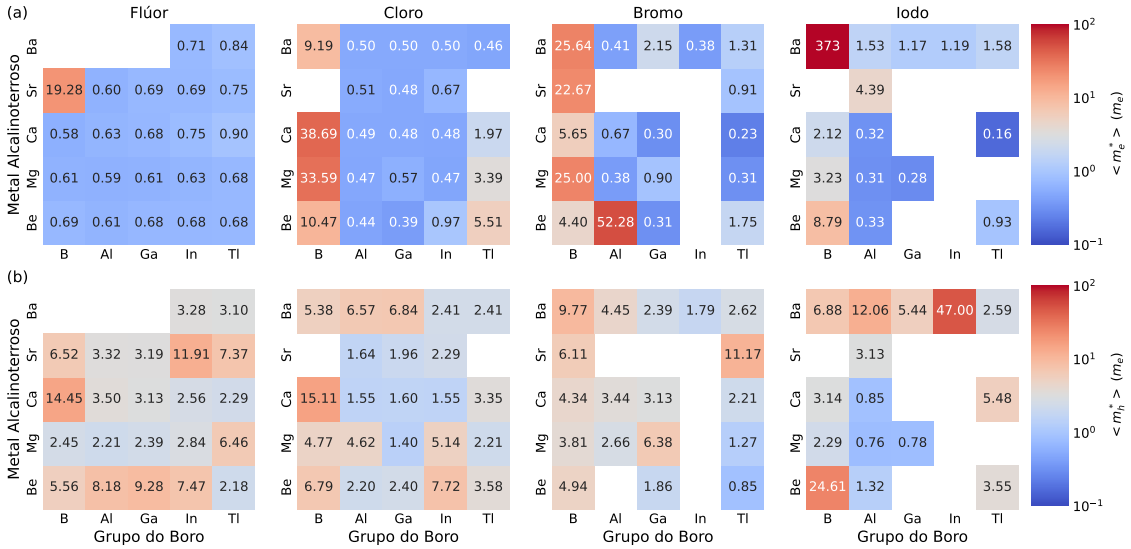
Figura 5.14: Estrutura eletrônica e densidade de estados (DOS) para $CaAlCl_5$, usando PBE-GGA. Gap de energia direto em Γ, $E_g = 3.6$ eV.



Fonte: elaborado pelo autor.

Através da estrutura de bandas conseguimos calcular a massa efetiva dos portadores de carga, sendo ela apresentada na Figura 5.15. O espaço vazio para $MgTlI_5$ corresponde ao fato desse material ser metálico ($E_g = 0$). Já o espaço vazio para $BeAlBr_5$ ocorre pois o resultado do ajuste parabólico da banda retorna um valor infinito - isso ocorre pois a banda está próxima da linha reta horizontal (assinatura conhecida em inglês como *flat band*), como pode ser verificado na respectiva estrutura de bandas no Apêndice B (Figura B18).

Figura 5.15: Massa efetiva dos portadores de carga: (a) elétrons, (b) buracos.



Fonte: elaborado pelo autor.

De maneira geral, os elétrons possuem as menores massas efetivas, exceto quando o material contém boro - neste caso os materiais apresentam massas efetivas do elétron bastante elevadas, como no caso do $BaBI_5$, com cerca de 373 vezes a massa de repouso do elétron.

5.2.4 Seleção

Com base nos resultados de estabilidade termodinâmica e estrutura eletrônica, podemos selecionar materiais candidatos para nossa aplicação alvo. Os requisitos de design são materiais estáveis ($\Delta H_f \leq 0$ e $E_h = 0$) com bandgap direto da ordem de 4.6 eV, com portadores de carga com massas efetivas pequenas.

Filtrando para $3 < E_g < 6$ com gap direto, sobraram 7 candidatos que são apresentados na Tabela 5.3. Nesta filtragem, somente $BeInF_5$ e $MgGaF_5$ estão no *convex hull*. Por sua vez, $oMgInF_5$ e o $CaInF_5$ apresentam distância muito pequena, apenas de 10 meV, o que normalmente não é um impediditvo forte para a observação desses materiais. Por outro lado, materiais com $E_h > 100$ meV podem ser menos prováveis de serem sintetizados, mas não devem ser descartados sumariamente tendo em vista que em condições de temperatura

e pressão finita, eles podem representar estados acessíveis na natureza e em laboratório.

Tabela 5.3: Resultados para o composto alvo $\alpha-CaAlF_5$.

Material	Estabilidade Termodinâmica			Propriedades Eletrônicas		
	E_t (eV)	ΔH_f (eV/atom)	E_h (eV/atom)	E_g (eV)	m_e^* (m_e)	m_h^* (m_e)
$MgAlCl_5$	−49.65	−1.77	0.19	3.36	0.47	4.62
$CaAlCl_5$	−54.80	−2.05	0.13	3.60	0.49	1.55
$SrAlCl_5$	−54.98	−2.13	0.12	3.71	0.51	1.64
$MgInF_5$	−66.51	−3.16	0.01	4.69	0.63	2.84
$CaInF_5$	−70.27	−3.34	0.01	5.08	0.75	2.56
$BeInF_5$	−69.86	−3.05	0.00	5.15	0.68	7.47
$MgGaF_5$	−68.40	−3.28	0.00	5.59	0.61	2.39

Fonte: elaborado pelo autor.

Para as propriedades elétricas, sabemos que o DFT PBE-GGA tende a subestimar o bandgap, de forma que valores menores que 4.6 eV podem render materiais aplicáveis para LEDs UV. Já a presença de excítons pode diminuir o gap ótico que o material apresenta. Podemos endereçar essa última questão com a mesma metodologia que utilizamos no Capítulo 4 para calcular a energia de ligação dos excítons de perovskitas através do modelo de Wannier.

Quanto à novidade desses compostos, não encontramos entradas prévias para os mesmos nas bases de dados utilizadas (Materials Project e OQMD) durante a execução desse trabalho. Na literatura, para o $MgAlCl_5$ encontramos apenas uma referência a sua formação em reações de catalise [291], sem detalhes sobre sua estrutura ou características; para $CaAlCl_5$ encontramos uma referência de 2010 que comenta que sua estrutura ainda não havia sido determinada [292] e uma referência de 2018 que atesta sua formação durante esfoliação química de silício [293]; Não encontramos referências diretas aos outros cinco materiais dessa lista.

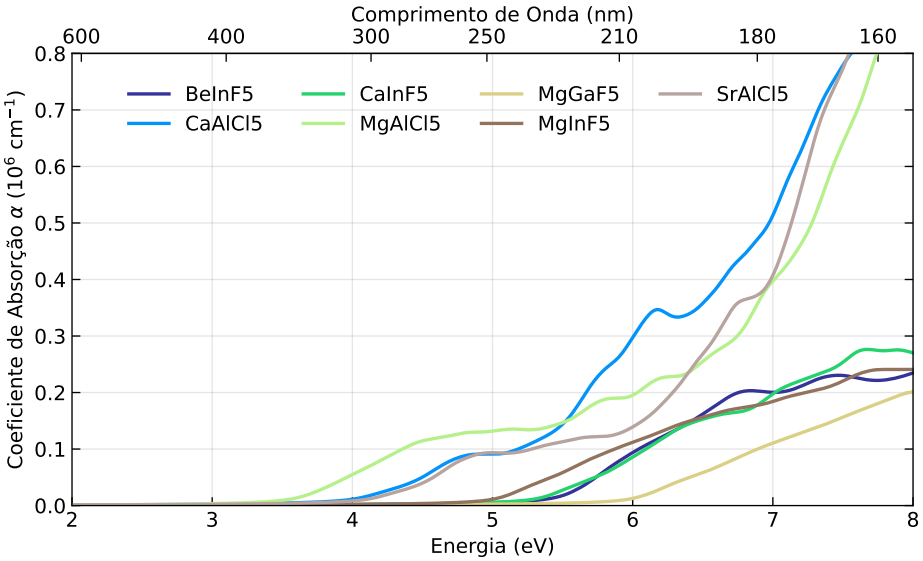
Devemos notar que boa parte das referências mais antigas para parte dessa família de materiais (como $CaAlF_5$ e $BaInF_5$) incluem vidros fluoretos [283] e vidros fluoroindatos [294]. Estes materiais estão em fase amorfa e são formados por misturas de composições. Os resultados reportados aqui são referentes a materiais cristalinos puros - não simulamos defeitos ou dopagem.

Baseado na heurística da nossa metodologia, as possíveis rotas de síntese para estes compostos selecionados são análogas as utilizadas para sintetizar a jakobssonita. A síntese de estado sólido pelo método de fluxo proposto por A. Hemon e G. Courbion [284] é uma dessas rotas, além do método sol-gel de Rehmer *et al.*, também já citado na introdução [278]. Weil *et al.* descrevem uma rota solvotermal para crescer $CaAlF_5$, $BaAlF_5$ e $BaAlF_5$ em autoclave a 250°C por 16 dias [295]. Novamente, lembramos que a transferibilidade dessas rotas não é garantida, mas constitui uma direção de investigação possível e promissora.

5.2.5 Propriedades Óticas

Para investigar a questão do gap ótico esperado, calculamos o espectros de absorção para os 7 materiais selecionados na etapa anterior - seguimos a mesma metodologia reporta no Capítulo 4. Esses resultados são apresentados na Figura 5.16.

Figura 5.16: Espectro de absorção dos materiais alvo selecionados, representado pelo coeficiente α obtido por DFT.



Fonte: elaborado pelo autor.

Utilizando a parte real da permissividade elétrica, calculamos a constante dielétrica isotrópica, tomando a situação de baixa frequência. Esses valores foram utilizados no modelo de Wannier (Equações 4.19 e 4.20) para calcular a energia de ligação do excíton E_b e o raio do excíton a_{ex}. Também utilizamos a Eq. 4.21 para calcular o campo elétrico de ionização do excíton. Sumarizamos esses resultados na Tabela 5.4.

Tabela 5.4: Constante dielétrica e propriedades do excíton para os materiais selecionados.

Material	ϵ_{iso}	μ_{ex} (m_e)	E_b (eV)	a_{ex} (Å)	$\lvert\vec{E}_c\rvert$ (10^9 V/m)
$MgAlCl_5$	3.604	0.427	0.447	4.47	2.00
$CaAlCl_5$	3.352	0.372	0.451	4.76	1.89
$SrAlCl_5$	3.193	0.389	0.519	4.34	2.39
$MgInF_5$	2.206	0.516	1.442	2.26	12.73
$CaInF_5$	2.246	0.580	1.564	2.05	15.27
$BeInF_5$	2.214	0.623	1.730	1.88	18.41
$MgGaF_5$	2.176	0.486	1.396	2.37	11.79

Fonte: elaborado pelo autor.

Dada as energias de ligação encontradas, o gap ótico esperado para esses materiais pode estar até 1.7 eV deslocado do gap eletrônico (caso do $BeInF_5$). Contudo, é importante lembrar que o modelo de Wannier é uma aproximação útil para excítons livres. Quando os raios de excíton encontrados são muito pequenos, da ordem das distâncias atômicas, a literatura sugere que o modelo de Wannier falha - os excítons passam a ser chamados de excítons de Frenkel nestes casos [14]. É exatamente a situação que vemos aqui, por exemplo, para o $MgAlCl_5$, a distância entre o Mg e o Al é de 4.6 Å e a distância entre o F e o Al é de 2.51 Å, enquanto o raio do excíton é de 4.47 Å.

Como esses materiais tem quantidade de átomos muito menor do que no caso estudado das perovskitas pseudo-2D (Capítulo 4) - todas as estruturas da família da jakobssonita, grupo alvo (SG = 15), têm 14 átomos no POSCAR, enquanto que para as perovskitas $n = 1$ há 62 átomos, $n = 2$, 172 átomos, e para a $n = 3$, 220 átomos - acreditamos que uma linha de investigação possível envolveria o uso de DFT corrigida por GW e cálculo BSE para aprimorar os resultados eletrônicos e óticos aqui discutidos. Isso permitiria ter um melhor modelo para lidar com os excítons além de corrigir a subestimação do bandgap.

5.3 Conclusão

Neste capítulo, realizamos cálculos de DFT *high throughput* para criar uma nova família de 100 materiais baseados na jakobssonita, constituídos pela combinação química dos metais alcalinoterrosos, os elementos do grupo do boro e os halogênios. Dividimos o problema do design de um semicondutor UWBG em duas frentes: a *sintetizabilidade* e as propriedades de interesse. Para resolver parte da *sintetizabilidade* utilizamos o critério de estabilidade termodinâmica que incluiu a simulação de 2833 outros materiais. Como propriedade de interesse, estudamos a estrutura eletrônica e o espectro de absorção, tentando determinar o gap eletrônico e o gap ótico dos materiais alvo. Como resultado, selecionamos sete materiais ($MgAlCl_5$, $CaAlCl_5$,$SrAlCl_5$,$MgInF_5$,$CaInF_5$,$BeInF_5$ e $MgGaF_5$) como semicondutores UWBG candidatos a aplicação em LEDs UV.

Para os materiais ternários simulados, tanto os alvos, quanto os secundários, verificou-se que o bandgap é controlado majoritariamente pelo halogênio: presença de flúor está relacionada a um maior bandgap, seguido do cloro, bromo e iodo. Para os materiais alvo selecionados, os raios de excíton calculados indicam que o uso do modelo de Wannier é limitado.

Do ponto de vista metodológico, neste estudo de caso expandimos a largura da busca dos materiais se comparada a quantidade de materiais estudados no estudo de caso I (Capítulo 4). De fato, nossos experimentos *in silico* puderam explorar não só os materiais alvo, mas também uma grande quantidade de fases competidoras que, renderam pelo menos 259 materiais estáveis, sendo 176 semicondutores. Dessa forma, o uso de HT-DFT gerou coprodutos que poderão ser utilizados em estudos futuros. Fica patente que, um dos resultados obtidos é exatamente a criação dessa nova base de dados de materiais computacionais com 2933 entradas. Infelizmente não temos espaço suficiente neste trabalho para

explorar todas as implicações desses resultados secundários, nos limitando as estatísticas gerais que discutimos neste capítulo e aos apêndices. Por outro lado, o fato de usar DFT de baixa fidelidade (PBE-GGA) para ter velocidade na exploração desse espaço, pode limitar a profundidade dos resultados finais. Podemos entender isso como uma imagem de um mapa: quanto mais área queremos cobrir, menor a definição que teremos do terreno. Para contornar isso, precisamos aprofundar, isto é, *dar zoom*, em poucos materiais - como fizemos nos sete selecionados -, para conseguir entende-los em mais detalhes. Neste caso, ainda estaremos numa metodologia de experimentação *in silico*, porém agora explorando *em profundidade* de vez de *em largura*.

Em relação a família de materiais da jakobssonita que estudamos aqui, visamos como próximas etapas o aprofundamento das simulações computacionais sobre os materiais selecionados: correção de gap com GW; cálculo de excítons com BSE; cálculo da estrutura de fônons para determinar estabilidade dinâmica; e cálculo de dinâmica molecular *ab initio* para verificar efeitos de temperatura e pressão na estabilidade do material. Para os materiais que forem selecionados por essa etapa, a última etapa é a síntese e caracterização.

A Ciência de Dados de Materiais, mais fortemente caracterizada neste capítulo pela sua vertente em Informática de Materiais (bases de dados e cálculos *high throughput*) também fornece outros meios para explorar em largura o espaço de materiais. Discutimos algumas dessas técnicas no estudo de caso III, relatado no próximo capítulo.

6 Estudo de Caso III: Semicondutores Binários

Desde a invenção do transistor de germânio por John Bardeen, Walter Brattain e William Shockley em 1947 [296], o interesse em materiais cristalinos semicondutores cresceu significativamente, impulsionando avanços na eletrônica, computação e nas tecnologias de comunicação. O germânio empregado nesse dispositivo apresenta um gap de energia pequeno, de aproximadamente 0.66 eV, um pouco mais da metade do gap de energia do silício, que é de cerca de 1.12 eV. Essa propriedade é essencial para o funcionamento do transistor, que hoje é, geralmente, construído com silício cristalino dopado. Da mesma forma, dispositivos LED necessitam de materiais com gaps de energia específicos, com cada cor de emissão representando uma faixa de valores: diodos emissores de luz vermelha são fabricados com materiais semicondutores como arseneto de gálio-alumínio (AlGaAs), arseneto-fosfeto de gálio (GaAsP) e fosfeto de gálio (GaP); já para a emissão na faixa do verde são empregados fosfeto de gálio (GaP), fosfeto de gálio-alumínio-índio (AlGaInP), nitrato de gálio (GaN) e nitrato de índio-gálio (InGaN), este último também possibilitando a construção de LEDs azuis, graças ao avanço tecnológico promovido por Shuji Nakamura em 1995 [297]; na região do ultravioleta, os LEDs utilizam materiais de gap ainda maior, como nitreto de alumínio (AlN), nitreto de boro (BN), diamante e ligas como alumínio-gálio-nitreto (AlGaN) ou alumínio-gálio-índio-nitreto (AlGaInN).

O atual interesse em materiais com gap de energia largo e ultralargo reside na possibilidade de aperfeiçoar as tecnologias vigentes e permitir novas tecnologias, em especial as tecnologias de eletrônica de potência, de eletrônica de comutação para radio frequência, tecnologias optoeletrônicas de ultravioleta profundo, e aplicações em informação quântica [298]. Por exemplo, a medida que os transistores ficam cada vez menores, as camadas isolantes entre os contatos de porta devem resistir a campos elétricos cada vez mais intensos, isto é, não conduzir eletricidade (corrente de fuga) - o que é muito relevante em transistores usados em radio-frequência. Da mesma forma, como discutimos acima para LEDs, gerar emissões em espectros de alta energia como o UV necessitam de materiais com gaps cada vez maiores.

Em vista de sistematizar o conhecimento produzido e organizar os materiais para cada aplicação, podemos introduzir uma classificação artificial dos materiais semicondutores com relação aos valores de bandgap. Na literatura atual [298, 299, 300] são empregados os termos semicondutores de gap largo (do inglês, *wide bandgap semicondutors*) e de gap ultralargo (*ultra-wide bandgap*) para se referir a materiais com grandes bandas proibidas - exemplos típicos de materiais WBGs são o GaN (3.4 eV) e o SiC (3.2 eV), enquanto de UWGBs são as ligas de $Al_xGa_{1-x}N$ (3.4-6 eV), o diamante (5.5 eV), o BN hexagonal (6

eV), os óxidos binários β-Ga_2O_3 (4.5 eV), α-Ga_2O_3 (5.3 eV), rutilo de GeO_2 (4.7 eV) e rutilo de SnO_2, e os óxidos ternários, espinela de $ZnGa_2O_4$, espinela de $MgGa_2O_4$, $SrTiO_3$, $BaSnO_3$ e $SrSnO_3$ [299, 300]. Em contrapartida, a literatura [301, 302] também apresenta o termo bandgap estreito (*narrow bandgap*) para se referir a materiais cuja banda proibida é pequena - exemplos incluem InSb (180 meV), InAs (356 meV) e GaSb (725 meV) [301]. Contudo, ainda não existe um consenso sobre limiares de valores que delimitam as faixas de classificação. Em vista de tornarmos o presente texto mais preciso no uso de tais termos, propomos na Tabela 6.1 uma sumarização de tais faixas com suas respectivas nomenclaturas.

Tabela 6.1: Nomenclaturas para classificação de bandgap.

Nome da Faixa	Sigla	Bandgap (eV)
Narrow Bandgap	NBG	$0 < E_g \leq 0.7$
Conventional Bandgap	CBG	$0.7 < E_g \leq 2$
Wide Bandgap	WBG	$2 < E_g \leq 3.6$
Ultra-wide Bandgap	UWBG	$3.6 < E_g$

Fonte: elaborado pelo autor.

Essa sistematização não apenas facilita a categorização dos materiais semicondutores existentes, como também se mostra promissora para aplicações algorítmicas, onde regras bem definidas permitem estabelecer os próprios métodos computacionais. Dessa vez, em contraste com os Capítulos 4 e 5, pretendemos demonstrar a aplicação de técnicas de IA preditiva para a geração de novos materiais partindo de uma abordagem puramente computacional, onde os princípios físicos são inseridos de maneira *ad hoc*. Como motivação, mantemos o interesse em semicondutores, visando a aplicação em LEDs emissores no ultravioleta.

Aqui gostaríamos de realizar uma inversão da ordem lógica apresentada nos métodos anteriores. Para se estudar computacionalmente e experimentalmente um material - como as pervoskitas híbridas do Capítulo 4 - partimos do material para as propriedades, ou seja, o bandgap é determinado após a definição precisa do material (conhecemos a composição química e a estrutura cristalina). Para jakbossonitas do Capítulo 5, também partimos de um material protótipo obtendo o bandgap apenas ao final do processo, porém expandimos o espaço explorado dessa propriedade ao criarmos uma família de novos materiais com estruturas bem definidas. Um método puramente inverso deveria ser capaz de partir do espaço total dessa propriedade para então definir o material específico que apresenta o valor de bandgap desejado. Em outras palavras, gostaríamos de gerar todos os valores de bandgap possíveis e então selecionar ou reconstruir um material que apresentasse determinado valor de gap. Porém, tal possibilidade seria estranha, já que o valor do gap de energia não é unívoco - faltariam informações para definir cada material univocamente no espaço de bandagap (precisaríamos, talvez, de um espaço definido sobre mais propriedades do

material). Outra possibilidade seria explorar todo o espaço de materiais de maneira mais eficiente. Isso é possível de ser feito com modelos de aprendizado de máquina devidamente treinados. Para tanto, precisamos representar o material em um outro espaço reduzido e usar um *surrogate model* para a predizer o bandgap da estrutura amostrada nesse espaço de representação.

Existem diferentes formas de alcançar esses objetivos com algoritmos de aprendizado de máquina, como discutido na Seção 3.3. Para os módulos preditores, usaremos algorítimos de aprendizado raso com uma representação que garanta certo nível de interpretabilidade e explicabilidade; já para o módulo de exploração do espaço representacional, usaremos combinatória por substituição química, como usado no Capítulo 5. Uma diferença importante é que naquele capítulo estávamos interessados em gerar somente alguns materiais possíveis, enquanto aqui iremos gerar todas as combinações permitidas pelo espaço representacional. Devido as essas duas condições impostas (interpretabilidade e exploração total do espaço), decidimos trabalhar apenas com materiais sólidos cristalinos binários. A junção desses dois módulos permitiu a construção de um sistema generativo que propõem novas estruturas para os valores de gap selecionados. É, dessa forma, um passo na direção de se ter um sistema de predição inverso.

Destacamos, portanto, que, neste capítulo, nosso interesse está nos métodos e seus resultados mais do que em um material específico. Podemos enxergar isso da seguinte forma: com um método generativo não estamos olhando apenas para um material, mas para toda uma classe de materiais que podemos efetivamente mapear com tais métodos.

Nossa abordagem constitui de duas etapas: na primeira utilizamos técnicas de ciência de dados para sistematizar o estudo desses materiais semicondutores; na segunda, usamos esse conhecimento para construir o modelo generativo de materiais que explora eficientemente o espaço representacional, permitindo mapear a propriedade de interesse para um grande volume de materiais protótipos

6.1 Metodologia

6.1.1 Datasets

Obtivemos dados sobre sólidos inorgânicos binários do Materials Project[146] (MP), totalizando 16775 entradas únicas (mp_id único). Esses dados foram acrescidos de informações atômicas oriundas do mendeleev [303]. Denominamos esse *dataset* composto de informações do MP com informações do mendeleev de MP-b0.

Com objetivo de cobrir as variações arbitrárias de escolha da ordem de uma composição, realizamos uma etapa de *data augmentation* no *dataset* MP-b0. A ideia baseia-se no fato de que, para os algoritmos de aprendizado de máquina usados, as predições realizadas para, por exemplo, NaCl devem ser as mesmas que para ClNa. Como será discutido na Seção 6.1.2, existem outras maneiras de considerar essa permutação, que dependem da escolha dos algoritmos a serem utilizados. Para nossos propósitos, o procedimento de *data*

augmentation realizado aumentou (dobrou) o número de instâncias (linhas), sem modificar o número de atributos (colunas) do *dataset*. Esse *dataset* aumentado denominamos MP-bA.

O uso do *dataset* MP-bA objetivou o treino e teste dos algoritmos de aprendizado de máquina. Por outro lado, o *dataset* aqui denominado GCC-0, constituiu na geração de novos compostos baseados em combinações dos elementos químicos e sistemas cristalinos. Esse *dataset* foi gerado sinteticamente por combinatória, de maneira analoga aquela proposta no Capítulo 5, com o número de compostos Ω podendo ser calculado por:

$$\Omega = N_p N_{sc}[N_{el}(N_{el} - 1)], \tag{6.1}$$

onde N_p é o número de diferentes proporções entre os átomos, N_{el} é o número de elementos únicos utilizados e N_{sc} é o número de sistemas cristalinos.

Na Tabela 6.2 estão as respectivas quantidades de entradas dos três *datasets* usados:

Tabela 6.2: Informações gerais dos *datasets* usados neste estudo de caso.

Nome do *dataset*	Definição	Quantidade de entradas
MP-b0	Dados de DFT oriundos do Materials Project, acrescidos de informações atômicas do Mendeleev	16775
MP-bA	Dados do MP-b0 aumentados por permutação das features atômicas	33550
GCC-0	Dados gerados por combinatória dos elementos químicos, estequiometrias e sistemas cristalinos	1380288

Fonte: elaborado pelo autor.

No nosso caso: $N_{el} = 79$, $N_{sc} = 7$ e $N_p = 32$. Ou seja, estamos combinando 79 elementos químicos, em 7 sistemas cristalinos diferentes, com 32 proporções diferentes. Nota-se que na Eq. 6.1, o termo $N_{el}(N_{el}-1)$ é a permutação dos elementos. Já N_p depende das quantidade únicas x e y de cada elemento. Se definirmos os conjuntos de valores de quantidade para cada elemento como $X = \{x_1, x_2, x_3, ..., x_n\}$ e $Y = \{y_1, y_2, y_3, ..., y_n\}$, então N_p estará contido entre $1 \leq \left|\left\{\frac{\min(x,y)}{\min(x,y)} \;\middle|\; x \in X,\, y \in Y,\, x \neq 0, y \neq 0\right\}\right| \leq n^2$. Com $n = 10$ valores diferentes para cada átomo, teríamos um limite superior de 100 combinações a serem exploradas, gerando 4313400 compostos. Contudo, tais combinação acabam por gerar repetições de proporção devido a estequiometrias equivalentes. Por estequiometria equivalente nos referimos a situação na qual um composto A_xB_y possui $x = y$ (A_1B_1, A_2B_2, A_3B_3 e assim por diante) ou $x \bmod y = 0 \vee y \bmod x = 0$ (A_1B_2, A_2B_4, A_1B_3, A_3B_9 e assim por diante). Queremos contar pelo menos uma das ocorrências das equivalências. Isso é o mesmo que contar a quantidade de frações irredutíveis no intervalo $[1, 10]$. São elas apresentadas na Tabela 6.3, conjuntamente com a fórmula genérica do composto e

a proporção calculada. Dessa forma, a quantidade de materiais gerados foi de 1380288 compostos químicos únicos.

Tabela 6.3: Proporções assumidas entre a quantidade de átomos de um composto binário.

Fração Irredutível	Proporção	Fórmula
1/1	100.0%	A_1B_1
1/2	50.0%	A_1B_2
1/3	33.3%	A_1B_3
1/4	25.0%	A_1B_4
1/5	20.0%	A_1B_5
1/6	16.6%	A_1B_6
1/7	14.3%	A_1B_7
1/8	12.5%	A_1B_8
1/9	11.1%	A_1B_9
1/10	10.0%	A_1B_{10}
2/3	66.6%	A_2B_3
2/5	40.0%	A_2B_5
2/7	28.6%	A_2B_7
2/9	22.2%	A_2B_9
3/4	75.0%	A_3B_4
3/5	60.0%	A_3B_5
3/7	42.9%	A_3B_7
3/8	37.5%	A_3B_8
3/10	30.0%	A_3B_{10}
4/5	80.0%	A_4B_5
4/7	57.1%	A_4B_7
4/9	44.4%	A_4B_9
5/6	83.3%	A_5B_6
5/7	71.4%	A_5B_7
5/8	62.5%	A_5B_8
5/9	55.5%	A_5B_9
6/7	85.7%	A_6B_7
7/8	87.5%	A_7B_8
7/9	77.7%	A_7B_9
7/10	70.0%	A_7B_{10}
8/9	88.8%	A_8B_9
9/10	90.0%	A_9B_{10}

Fonte: elaborado pelo autor.

6.1.2 Aprendizado de Máquina

6.1.2.1 Algoritmos

Escolhemos utilizar majoritariamente algoritmos de aprendizado raso presentes na biblioteca Scikit-Learn [304], para as tarefas de regressão e classificação. Os algoritmos utilizados são listados na Tabela 6.4:

Tabela 6.4: Algoritmos de ML utilizados neste trabalho.

Tipo	Classificação	Regressão
Modelos Lineares	LogisticRegression LogisticRegression Lasso LogisticRegression Ridge LogisticRegression Elastic	LinearRegression (OLS) Lasso Ridge ElasticNet
Modelos de Vizinhos	KNeighborsClassifier	KNeighborsRegression
Modelos de Árvores	DecisionTreeClassifier	DecisionTreeRegressor
Modelos de Ensemble	RandomForestClassifier GradientBoostingClassifier	RandomForestRegressor GradientBoostingRegressor
Modelos Neurais	MLPC	MLPR

Fonte: elaborado pelo autor.

Exceto pelo algoritmos de Regressão Logística no qual variamos o hiperparâmetro de penalidade para corresponder as versões Lasso, Ridge e Elastic, não realizamos nenhum ajuste de hiperparâmetros, utilizando os valores *default* sugeridos pela biblioteca. Evidentemente isso é uma limitação metodológica para que pudéssemos dar enfoque a outras variações, constituindo portanto, uma linha de investigação para trabalhos futuros. Por exemplo, o uso de *GridSearch* e *RandomSearch* para otimização dos hiperparâmetros é prevista na própria biblioteca do Scikit-Learn.

6.1.2.2 Features

O uso de algoritmos de aprendizado raso evidencia a necessidade de uma etapa de Engenharia de Features, na qual decidimos pelo uso de *features* atômicas para representar os materiais. Os valores para cada uma dessas *features* foram obtidos na ferramenta Mendeleev [303]. Como utilizamos um *dataset* de compostos binários, para cada elemento composto temos uma 11-tupla de *features* atômicas (Tabela 6.5), de forma que a representação total é a concatenação da tupla do primeiro elemento com a tupla do segundo elemento. As estequiometrias foram representadas por uma 2-tupla, onde a primeira posição representa a quantidade de átomos do primeiro elemento e a segunda, do segundo elemento. Além disso, temos características do composto, como o sistema cristalino que

foi representado como uma 7-tupla através da técnica de *One-Hot Encoding* (OHE). As *features* estruturais usadas são apresentadas na Tabela 6.6. Outras variáveis do material são o bandgap (E_g), a entalpia de formação (ΔH_f) e a distância para o *convex hull* (E_h). Essas informações, assim como sistema cristalino e quantidade de átomos de cada elemento, foram obtidas no Materials Project [146].

Tabela 6.5: *Features* atômicas utilizados neste trabalho.

Feature	Descrição
`period`	período na tabela periódica
`atomic_number`	número atômico
`atomic_radius`	raio atômico
`nvalence`	número de elétrons de valência
`dipole_polarizability`	polarizabilidade de dipolo
`electronegativity_pauling`	eletronegatividade de Pauling
`zeff`	carga nuclear efetiva
`s`,`p`,`d`,`f`	número total de elétrons nas camadas s, p, d e f

Fonte: elaborado pelo autor.

Tabela 6.6: *Features* estruturais utilizados neste trabalho.

Feature	Descrição
`nelement`	número de átomos daquele elemento
`Monoclinic`	se a estrutura do material é esse sistema vale 1, do contrário 0
`Tetragonal`	*id*
`Trigonal`	*id*
`Triclinic`	*id*
`Orthorhombic`	*id*
`Hexagonal`	*id*
`Cubic`	*id*

Fonte: elaborado pelo autor.

6.1.2.3 Avaliação de Desempenho

Utilizamos MAE, MAD, RMSE, STD, e $\bar{R}^2$ como métricas de desempenho de regressão. Para classificação utilizamos Acurácia, Precisão, Revocação, F1-score, AUC e coeficiente Kappa de Cohen.

Para entender questões de variabilidade intrínsecas ao *dataset*, utilizamos validação

cruzada sobre todo o conjunto. As divisões corresponderam a 5 conjuntos (denominados *folds* no jargão de ML), de maneira que foram utilizados 80% dos dados para treinamento e 20% para teste.

Aqui precisamos clarificar um detalhe: nas metodologias tradicionais de ML, o *dataset* é dividido em treino e teste e depois a porção de treino é novamente dividida entre treino e validação. A validação exerce o papel de teste nessa etapa e é usada para selecionar o melhor modelo (algoritmo e hiperparâmetros). No nosso caso utilizamos a validação cruzada sobre o *dataset* observando a variabilidade entre modelos e entre dados selecionados. O objetivo era verificar se uma mudança aleatória no conjunto de treinamento poderia causar um impacto muito grande no desempenho do mesmo algoritmo. Isso poderia acontecer, por exemplo, se treinássemos somente com semicondutores de gap pequeno e testássemos com semicondutores de gap grande. Fazendo uma amostragem aleatória, visamos manter a mesma distribuição nos dois conjuntos de forma que eventuais OOD (dados fora da distribuição, do inglês, *out-of-distribution*) correspondessem realmente às limitações de generalização do modelo.

A variabilidade que esses conjuntos causam sobre o desempenho dos modelo preditivos foi calculada pelo desvio padrão entre os testes de validação cruzada. Nos gráficos, optamos por representar os resultados pela média dos valores e as barras de erro pela a incerteza da média, tendo esta sido calculada pela divisão do desvio padrão pela raiz quadrada do número de testes realizados.

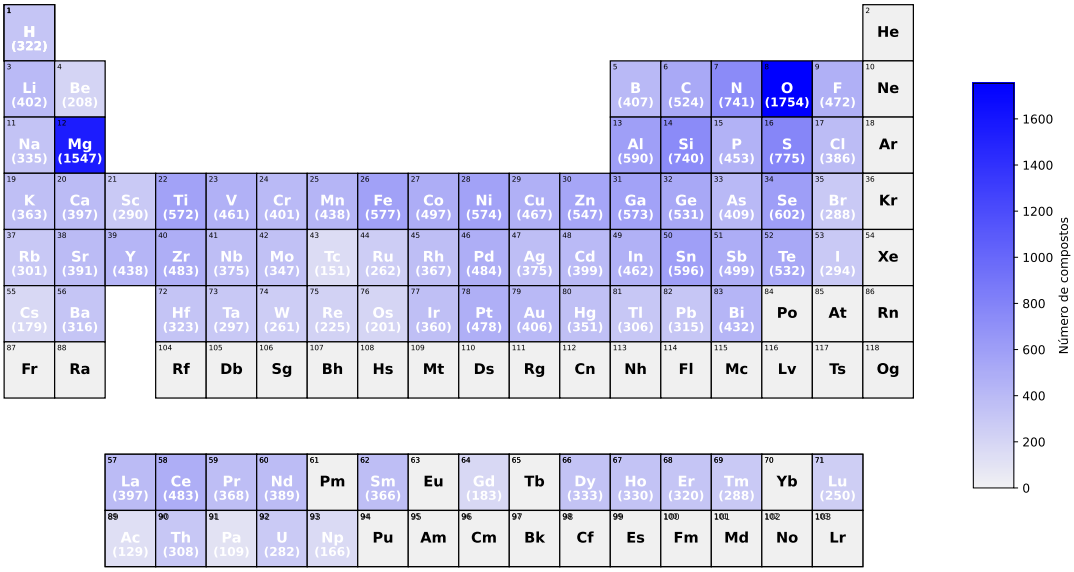
No caso do *dataset* MP-b0, a validação cruzada foi estratificada apenas com relação a metais e não metais, e/ou a estáveis e não estáveis. Para o caso do *dataset* MP-bA, precisamos estratificar a divisão de dados também com relação ao composto simétrico criado. Em outras palavras, precisamos garantir que as instâncias de NaCl e de ClNa estejam ambas ou no conjunto de teste ou no conjunto de treino para que de fato não haja vazamento de dados para o modelo.

6.2 Resultados e Discussão

6.2.1 Análise Exploratória

A análise exploratória visa investigar as estatísticas e comportamentos básicos dos dados utilizados. Iniciamos a discussão com a distribuição dos compostos binários estudados ao longo dos seus elementos atômicos constituintes, conforme Figura 6.1. Como pode ser observado, não são estudados compostos com gases nobres; a maioria dos elementos radioativos, em especial, os transurânicos sintéticos, também não constituem compostos da base; o elemento oxigênio (O) é o mais presente, seguido do magnésio (Mg). A prevalência de oxigênio pode ser entendida pela ampla presença de oxigênio na atmosfera terrestre, contudo, deve-se destacar que todas essas características levantadas são intrínsecas da base da dados (Materials Project), refletindo a escolha deliberada dos pesquisadores que simularam tais compostos, o que pode envolver uma gama diversa de interesses.

Figura 6.1: Distribuição dos elementos presentes nos materiais binários contidos no *dataset* MP-b0. O número de compostos é metade da soma de todas as contagens.



Fonte: elaborado pelo autor.

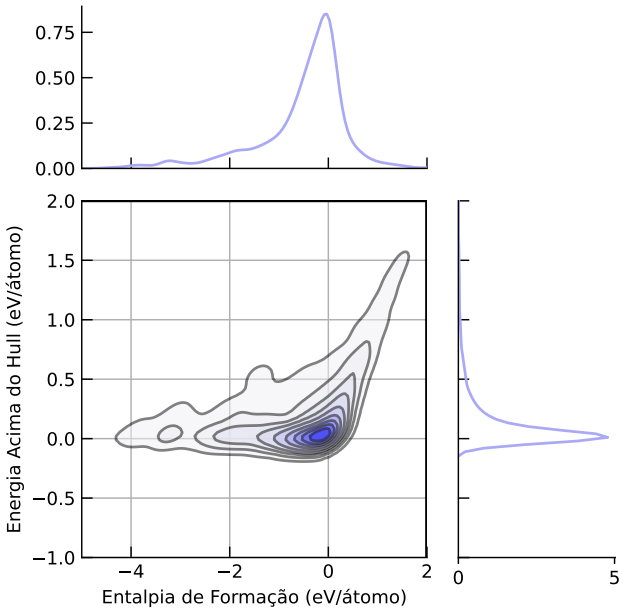
As quantidade de compostos presentes no *dataset* MP-b0 de acordo com essas propriedades podem ser vistas na Tabela 6.7 (para MP-bA a distribuição é exatamente a mesma, com valores absolutos multiplicados por dois). Observa-se que o *dataset* é composto em sua maior parte por metais (aproximadamente 80%) estáveis (aproximadamente 70%). Obviamente o critério de estabilidade aqui está relaxado, uma vez que estamos considerando apenas a decomposição com relação as fases elementares ($H_f \leq 0$). Poderíamos contabilizar com relação ao critério de distância para o *convex hull*, entretanto esse critério poderia dificultar as etapas subsequentes de aprendizado de máquina por dois motivos, um de natureza na Ciência de Dados e outro na Física de Materiais: respectivamente, a diminuição do total de dados disponível para treinamento dos modelos, e o fato de que o bandgap deve ser definido pela estrutura eletrônica intrínseca do material e não pela sua relação com as energias de outras fases de materiais com a mesma composição. Com relação ao segundo argumento, plotamos os gráficos das Figuras 6.2 e 6.4. No primeiro vemos que não há relação direta entre a entalpia de formação e a distância para o hull, quando consideramos materiais com entalpia de formação abaixo de zero; para valores acima do zero a relação é aproximadamente linear, o que era esperado já que estamos aumentando a energia livre associada à estrutura, o que, por sua vez, se reflete em uma maior distância ou energia necessária para mover o material para um estado mais estável (um ponto no hull). Na segunda Figura vemos que não há relação direta entre o bandgap e a distância para o hull no caso dos semicondutores.

Tabela 6.7: Contagem de compostos em MP-b0 separados por tipo de condutividade e por critério de estabilidade termodinâmica (todos possuem energia total menor do que zero).

Tipo de Condutividade	Critério de Estabilidade			Total
	$H_f > 0$	$H_f \leq 0$		
		$E_h > 0$	$E_h = 0$	
Metais $E_g = 0$	4340	5382	3579	13301
Semicondutores $E_g > 0$	307	2084	1083	3474

Fonte: elaborado pelo autor.

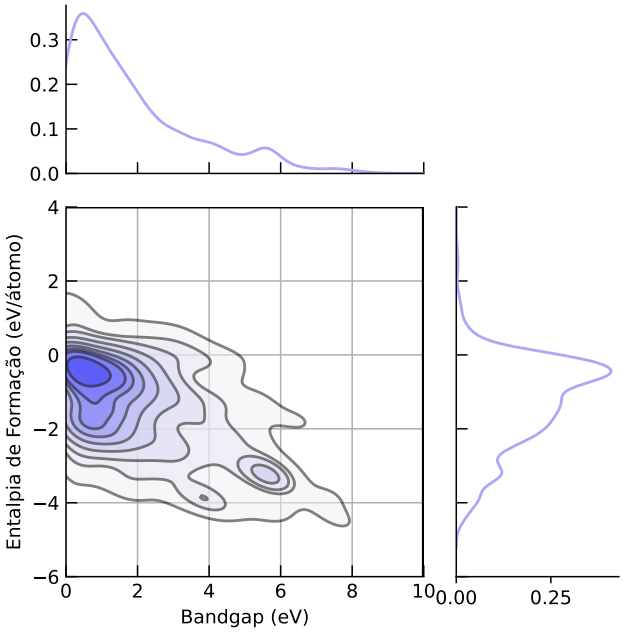
Figura 6.2: Distribuição dos valores de entalpia de formação e distância para o convex hull para o *dataset* MP-b0. Os valores assinalados nos gráficos laterais representam contagens relativas (frequências), assim como as curvas de nível no gráfico de dispersão.



Fonte: elaborado pelo autor.

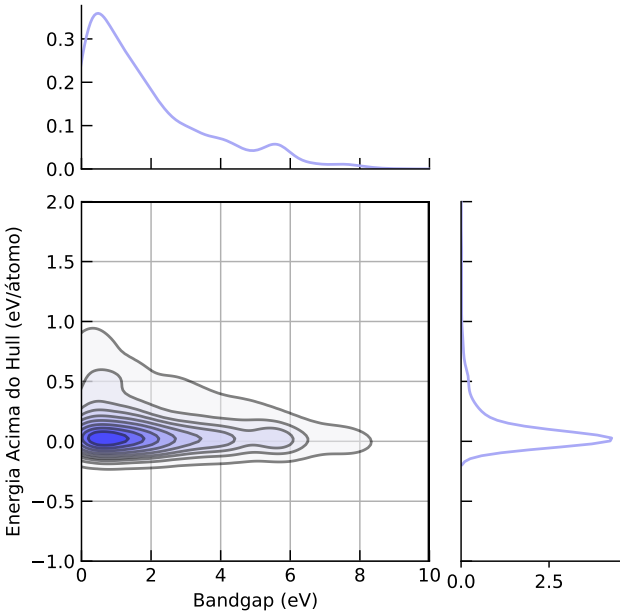
A Figura 6.3 representa as distribuições dos valores de entalpia de formação e energia de gap somente para os semicondutores. O objetivo da dispersão é entender como os materiais se distribuem ao longo de ambas as variáveis concomitantemente: há uma tendência de quanto maior o bandgap, menor a entalpia de formação; isso também corresponde a uma menor quantidade de dados, de forma que a maioria dos materiais estão concentrados com entalpia de formação próxima de zero e bandgap menor do que dois. Já a Figura 6.4 evidencia que, para este esquema de agrupamento, não há nenhuma relação de tendência entre o bandgap e a distância de energia para o *convex hull*, como discutimos anteriormente.

Figura 6.3: Distribuição dos semicondutores de acordo com a entalpia de formação e energia de bandgap para o *dataset* MP-b0. Os valores assinalados nos gráficos laterais representam contagens relativas (frequências), assim como as curvas de nível no gráfico de dispersão.



Fonte: elaborado pelo autor.

Figura 6.4: Distribuição dos semicondutores de acordo com a distância de energia para o *convex hull* e a energia de bandgap para o *dataset* MP-b0. Os valores assinalados nos gráficos laterais representam contagens relativas (frequências), assim como as curvas de nível no gráfico de dispersão.



Fonte: elaborado pelo autor.

As curvas de distribuição apresentadas nas Figuras 6.2, 6.3 e 6.4 são construídas pela técnica de KDE (do inglês, *Kernel Density Estimation*[1]) do próprio módulo de plote, sendo uma estimativa não paramétrica. Em vista de um maior rigor analítico, discutiremos uma verificação paramétrica por ajuste de dados.

Com esta finalidade, testamos as funções Normal, Lognormal, Gama, Exponencial, Weibull, Gumbel, Pareto e Rayleigh, Logística, t de Student e Cauchy para as distribuições de bandgap, entalpia de formação e energia acima do hull. Para determinar a aderência estatística utilizamos o teste de Kolmogorov–Smirnov (KS).

Em um primeiro momento, para todas as funções testadas, apenas a modelagem de bandgap com a distribuição de Weibull para os semicondutores rendeu um ajuste com p-valor acima de 0.05 (no caso, 0.0707 e distância KS de 0.0219). Os dados não fornecem evidência estatisticamente significativa para afirmar que as distribuições são diferentes (ao nível de 5%). Em outras palavras o p-valor não é alto o suficiente para dizer com confiança que as distribuições são exatamente iguais — só não há evidência forte o bastante para dizer que são diferentes. Portanto, aceitamos que os dados podem ter vindo da mesma distribuição, com base no teste KS.

A distribuição de Weibull é apresentada na Eq. 6.2. Com o ajuste realizado, obtivemos os seguintes valores de parâmetros: $k = 0.9601$, $\mu = 0.0003$ e $\lambda = 1.8188$.

$$f(x;\lambda,k,\mu) = \begin{cases} k\lambda\left[\lambda(x-\mu)\right]^{k-1} e^{-\left[\lambda(x-\mu)\right]^{k}}, & x \geq \mu \\ 0, & x < \mu \end{cases} \tag{6.2}$$

Nas Equações 6.2, 6.3 e 6.4, $k > 0$ é o parâmetro de forma, que controla o formato da distribuição. Quando $k < 1$, os valores de bandgap se concentram mais fortemente em valores baixos; se $k = 1$, a distribuição é assimétrica e decrescente, semelhante à exponencial; e se $k > 1$, a distribuição apresenta uma moda, com a maioria dos valores de bandgap próximos a um valor típico e caudas decrescentes. O parâmetro $\lambda > 0$ é o parâmetro de escala, que determina o espalhamento da distribuição, influenciando a largura da faixa de valores de bandgap observados. Já μ é o parâmetro de localização, que desloca a distribuição ao longo do eixo x, definindo o valor mínimo possível de bandgap considerado.

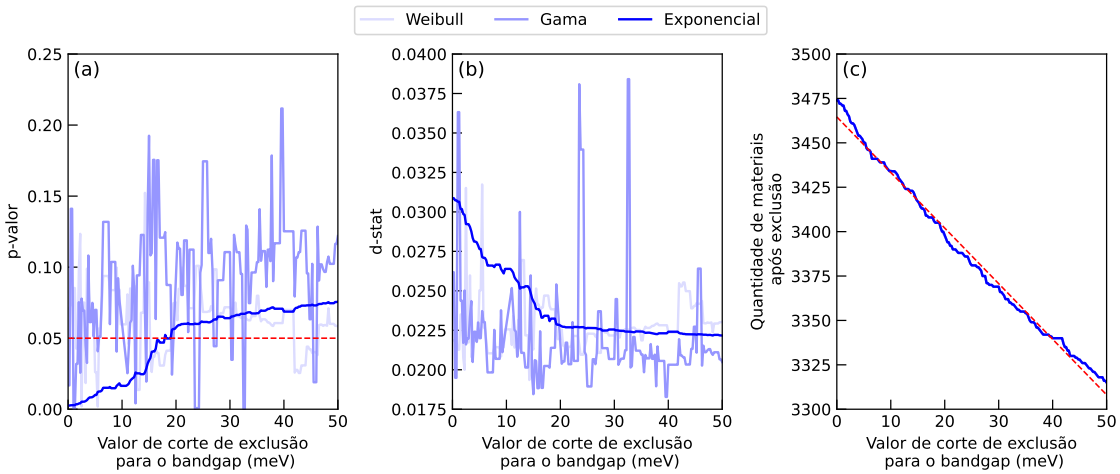
Essa análise nos indica que o bandgap mínimo considerado para início da distribuição é de exatamente $\mu = 0.0003$ eV, o que de fato podemos verificar ao procurar explicitamente esse valor nos dados. Porém, esse valor é muito pequeno e poderíamos argumentar plausivelmente que pode estar no limite de resolução numérica da DFT. Logo, precisamos estabelecer um segundo critério além dos limiares de separação (no caso em discussão, $E_g > 0$). Esse segundo critério não seria de fato um novo limiar de separação, mas sim um limiar de exclusão. Isto é, precisamos delimitar uma região entorno de $E_g = 0$ na qual não podemos afirmar que os materiais são semicondutores ou metais. Infelizmente, só podemos

[1]KDE no Matplotlib e Seaborn é uma técnica de suavização usada para estimar a distribuição de probabilidade de uma variável contínua, criando uma curva suave em vez de um histograma.

fazer isso para região de semicondutores, a qual temos os dados. Para os metais ($E_g = 0$) precisaríamos ter acesso a estrutura de bandas de cada material para poder estimar a largura de banda disponível de acordo com a Energia de Fermi - se o nível de Fermi corta a banda de condução em um determinado ponto, precisaríamos calcular quão distante ele está do último nível ocupado dentro da banda, de forma a estabelecer se este valor é maior do que o limite de resolução adotado.

Dessa forma variamos o valor de corte de exclusão para diferentes valores de bandgap verificando como isso impactava no ajuste das curvas de distribuição paramétrica. Pegando apenas as curvas com p-valor acima de 0.05, para um passo de energia de corte de 0.25 meV, obtivemos o comportamento apresentado na Figura 6.5:

Figura 6.5: Testes de distribuição contínua paramétrica para os semicondutores do *dataset* MP-b0. (a) p-valor do teste KS; a linha tracejada é o limiar p-valor=0.05. (b) d-stat para o teste KS. (c) Quantidade de dados mantidos; a linha tracejada representa o ajuste de tendência linear.



Fonte: elaborado pelo autor.

Nota-se que os p-valor para Weibull e Gama na Figura 6.5 (a) sofrem mudanças bruscas com o valor de corte considerado. A distância entre a distribuição empírica e a estimada pelo modelo paramétrico (d-stat) em (b) também tem o mesmo comportamento ruidoso. Por outro lado, a distribuição exponencial mostra-se suave nessa granularidade de resolução, com um crescimento constante em (a) e um decrescimento constante em (b). Em todos os valores de limiar testados, o decréscimo de dados em MP-b0 foi de no máximo 4.5% para 50 meV (de 3474 semicondutores para 3315), seguindo uma tendência aproximadamente linear no intervalo testado, como apresentado em (c).

Observando a natureza matemática das distribuições estudadas, temos que a distribuição Gama é dada por:

$$f(x;\lambda,k,\mu) = \begin{cases} \frac{\lambda^k (x-\mu)^{k-1}}{\Gamma(k)} e^{-\lambda(x-\mu)}, & x \geq \mu \\ 0, & x < \mu \end{cases} \tag{6.3}$$

Enquanto a distribuição exponencial é:

$$f(x;\lambda,k,\mu) = \begin{cases} \lambda e^{-\lambda(x-\mu)}, & x \geq \mu \\ 0, & x < \mu \end{cases} \tag{6.4}$$

Para o melhor ajuste da distribuição de Weibull (corte de 14.25 meV, p-valor de 0.15), obtivemos $k = 0.9733$, $\mu = 0.0143$ e $\lambda = 1.8686$; o melhor ajuste da distribuição Gama (corte de 39.50 e 39.75 meV, p-valor de 0.21), obtivemos $k = 0.9884$, $\mu = 0.0399$ e $\lambda = 1.9384$; para o melhor ajuste da distribuição exponencial (corte de 50 meV, p-valor de 0.07) obtivemos $\mu = 0.0503$ $\lambda = 1.8964$. Em todos os casos em que p-valor é maior do que 0.05, não podemos afirmar que a distribuição ajustada de fato é a distribuição dos dados, porém podemos afirmar que não podemos descartar estatisticamente que os dados observados não pertençam a tal distribuição. Por outro lado, os valores dos modelos obtidos sugerem que $k \approx 1$, enquanto que $\mu \approx 0$ - só é diferente de zero por causa do menor valor de bandgap presente nos dados (que, inclusive, é estabelecido pelo limite inferior do valor de corte de exclusão que estamos considerando). Assim, para $k \approx 1$ temos que:

$$f(x;\lambda,k,\mu) = k\lambda\left[\lambda(x-\mu)\right]^{k-1} e^{-[\lambda(x-\mu)]^k} = \lambda\left[\lambda(x-\mu)\right]^{1-1} e^{-[\lambda(x-\mu)]^1} = \lambda e^{-\lambda(x-\mu)}, \tag{6.5}$$

$$f(x;\lambda,k,\mu) = \frac{\lambda^k (x-\mu)^{k-1}}{\Gamma(k)} e^{-\lambda(x-\mu)} = \frac{\lambda^1 (x-\mu)^{1-1}}{\Gamma(1)} e^{-\lambda(x-\mu)} = \lambda e^{-\lambda(x-\mu)}. \tag{6.6}$$

Logo, dada a natureza ruidosa dos testes com Weibull e Gama, e das tendências que estamos observando nos valores dos modelos ajustados, é justificável supor que a distribuição dos valores segue uma distribuição exponencial, mesmo que o p-valor obtido tenha sido menor para este caso. Além disso, o valor de corte para a região de exclusão pode ser considerado próximo a 20 meV, devido a tendência apresentada na curva de p-valor e d-stat da distribuição exponencial - nota-se que, excluir materiais acima de 20 meV passa a não impactar o modelo da mesma forma que impactava ao excluir-se valores abaixo de 20 meV.

Uma vez que consideremos que o modelo obtido é uma extrapolação correta das observações realizadas, podemos afirmar que para o universo total de materiais binários semicondutores possíveis, a distribuição dos valores de gap segue uma curva de distribuição exponencial, com função densidade de probabilidade $f(E_g)$ dada por:

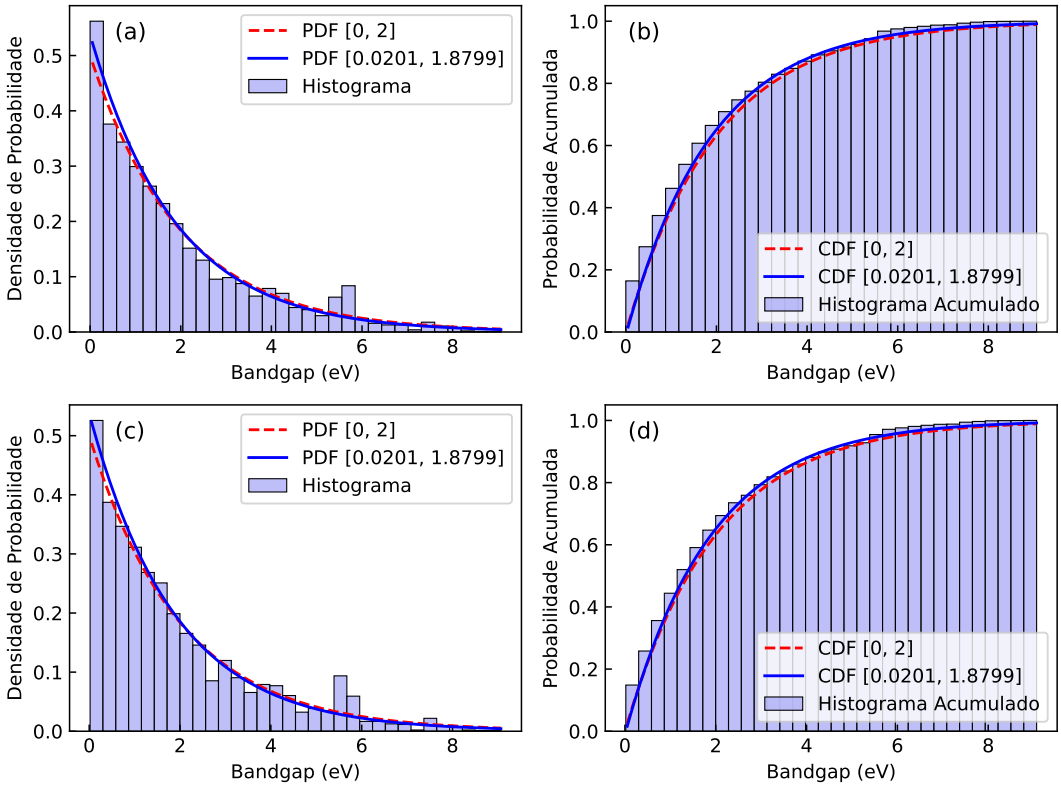
$$f(E_g) = 2e^{-2E_g}. \tag{6.7}$$

E a probabilidade $P(a \leq E_g \leq b)$ de um material binário ter um gap de energia na faixa entre a e b é de:

$$P(a \leq E_g \leq b) = \int_a^b 2e^{-2E_g} dE_g = e^{-2a} - e^{-2b}. \tag{6.8}$$

Há, contudo, uma pequena discrepância entre o ajuste e os dados observados próximo a 6 eV, como pode ser visto na Figura 6.6. Isso significa que há mais dados observados do que era esperado pelo modelo de distribuição de probabilidade para essa região. Isso pode ser um indicativo de que a curva ajustada não corresponde de fato a natureza de distribuição dos bandgaps ou que essa região está sobreamostrada com relação ao restante de dados. Neste último caso é esperado que, ao aumentar-se os dados correspondentes as outras regiões de bandgap, a protuberância próxima a 6 eV não seja mais vista.

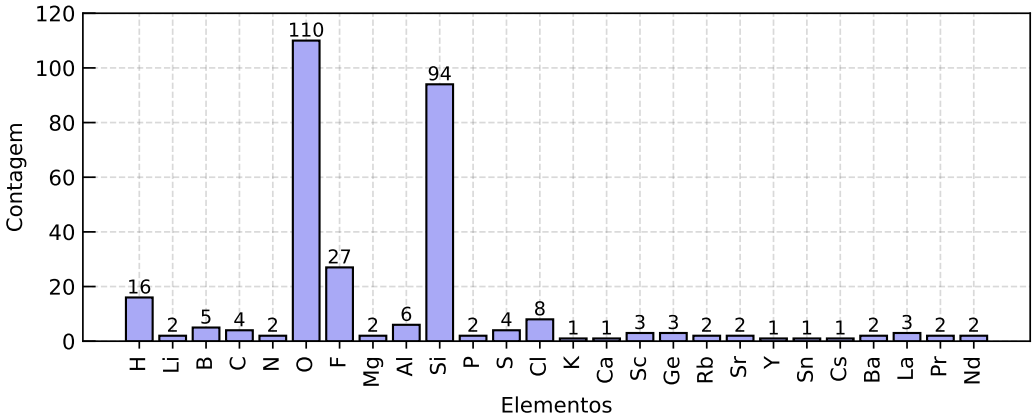
Figura 6.6: Distribuição do bandgap dos semicondutores para o *dataset* MP-b0. (a) Histograma e função densidade de probabilidade (PDF) exponencial ajustada para $E_g > 0$. (b) Histograma acumulado e função de distribuição acumulada (CDF) exponencial ajustada para $E_g > 0$. (c) Histograma e PDF exponencial ajustada para $E_g > 20$ meV. (b) Histograma acumulado e CDF exponencial ajustada para $E_g > 20$ meV. Para todos os histograma foram usados 31 bins de acordo com a regra de Rice.



Fonte: elaborado pelo autor.

Essa protuberância ocorre em um intervalo aproximado entre 5.4 e 6.1 eV. Se observamos a distribuição de elementos presentes nesses compostos binários na Figura 6.7, temos um forte indicativo de que se trata de uma sobreamostragem referente aos óxidos de silício, amplamente estudados devido a tecnologia vigente de semicondutores:

Figura 6.7: Distribuição dos elementos presentes nos materiais binários semicondutores para o *dataset* MP-b0 no intervalo de anomalia da distribuição de bandgap (entre 5.4 e 6.1 eV). O número de compostos é metade da soma de todas as contagens.

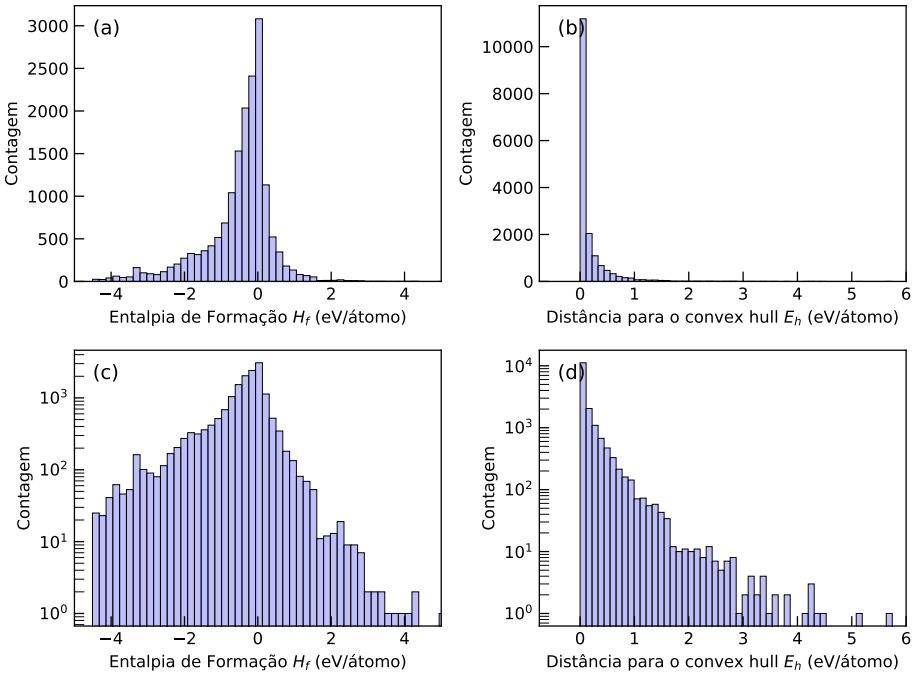


Fonte: elaborado pelo autor.

Para a análise paramétrica da distribuição da entalpia de formação H_f e da distância para *convex hull* E_h, construímos os gráficos da Figura 6.8. Tanto para H_f quanto para E_h nenhuma das distribuições testadas renderam p-valor maior do que zero. Para H_f os dados estão centrados em zero, com decaimento assimétrico em ambos os lados, conforme Figura 6.8 (a). Para E_h os dados seguem um decaimento rápido para valores positivos, conforme Figura 6.8 (b). Nestes casos, não é óbvia a necessidade de algum valor de corte, já que os valores naturalmente começam em zero estendo-se para qualquer valor positivo para todos os materiais - lembramos que pela construção da base de dados de origem (Materials Project), nenhuma valor negativo de E_h é mantido (um material mais estável do que todos os anteriormente simulados que compõem a superfície convexa), uma vez que a base dados é atualizada para recalcular a superfície, redefinindo o zero.

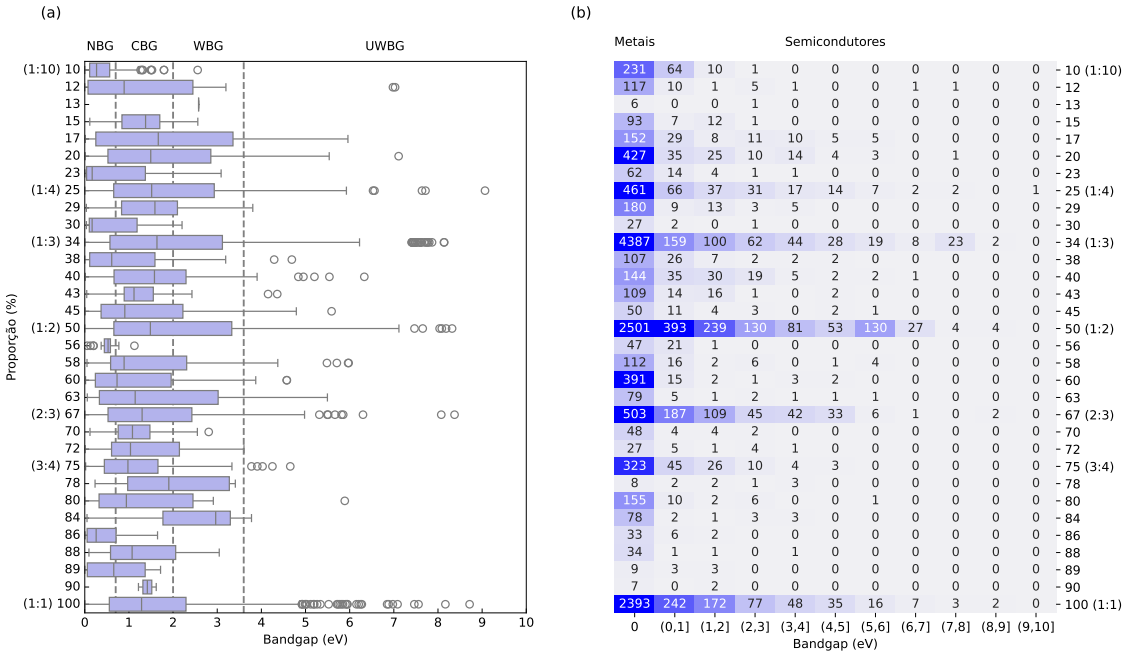
Neste primeiro desenvolvimento, estávamos tratando de três variáveis contínuas. A seguir trataremos das distribuições com relação a proporção entre os elementos de composição. Como neste caso temos dados discretizados em uma das variáveis (a proporção se reduz a 92 frações irredutíveis), decidimos utilizar uma outra forma de representar as distribuições. Na Figura 6.9 (a) estão presentes apenas os materiais semicondutores, com a distribuição dos gaps de energia dadas continuamente pela representação do boxplot ao longo do eixo x, e discretizado no eixo y pelas proporções assinaladas (reduzimos para 32 conforme valor definido para utilização na geração dos materiais, de maneira que tais valores definem, portanto, um intervalo). Nota-se que semicondutores UWBG sempre estão nos quartos quartis das distribuições ou configuram-se *outliers* para aquela proporção. Já em (b) estão inclusos os metais na primeira coluna da matriz. Aqui podemos entender que proporções arbitrárias entre os elementos dos compostos binários são tradicionalmente associadas a ligas cristalinas metálicas, enquanto que para cristais semicondutores, as proporções tendem a apresentar raias bem delimitadas em A_1B_4, A_1B_3, A_1B_2, A_2B_3 e A_1B_1.

Figura 6.8: Distribuição do entalpia de formação (a) e da distância para o *convex hull* (b) para *dataset* MP-b0. (c) e (d) correspondem respectivamente (a) e (b), porém em escala logarítmica no eixo y.



Fonte: elaborado pelo autor.

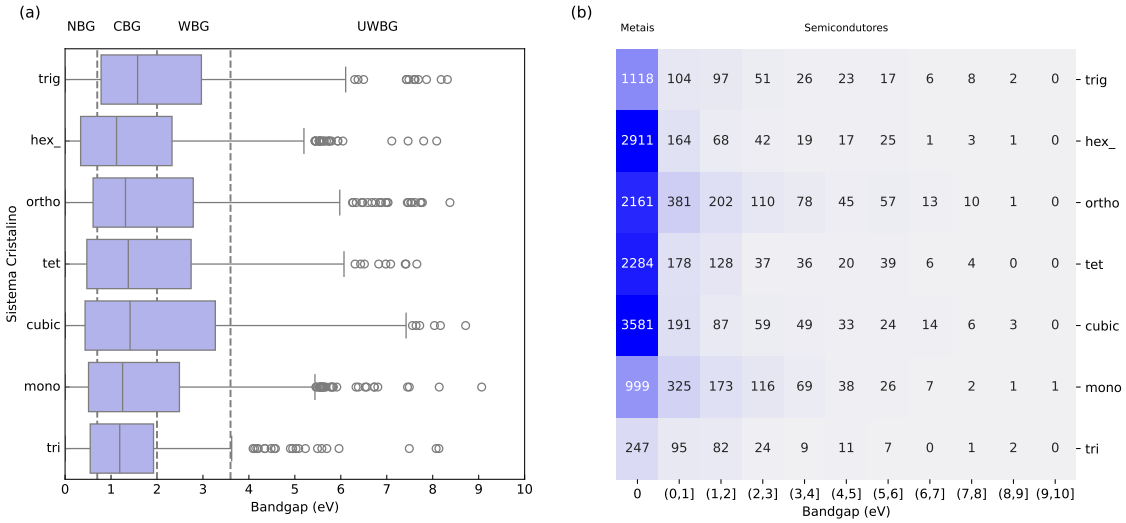
Figura 6.9: Distribuição das energias de gap em função da proporção entre os elementos para o *dataset* MP-b0. Em (a) não estão inclusos os metais.



Fonte: elaborado pelo autor.

A distribuição do bandgap com relação ao sistema cristalino é apresentada na Figura 6.10. Em (a) podemos observar que não há uma diferença significativa com relação ao grupo cristalino, exceto para os triclínicos que possuem menor quantidade de dados, sendo todos aqueles UWGB associados a *outliers* da distribuição. Em (b), diferentemente do agrupamento para proporção entre elementos, o agrupamento por sistema cristalino não possui raias bem definidas, com uma sutil predominância apenas para sistemas monoclínicos e ortorrômbicos ao longo dos gaps maiores. Observa-se também que há uma baixa quantidade de sistemas triclínicos, que corresponde a um sistema com pouca simetria ($a \neq b \neq c$ e $\alpha \neq \beta \neq \gamma \neq 90°$).

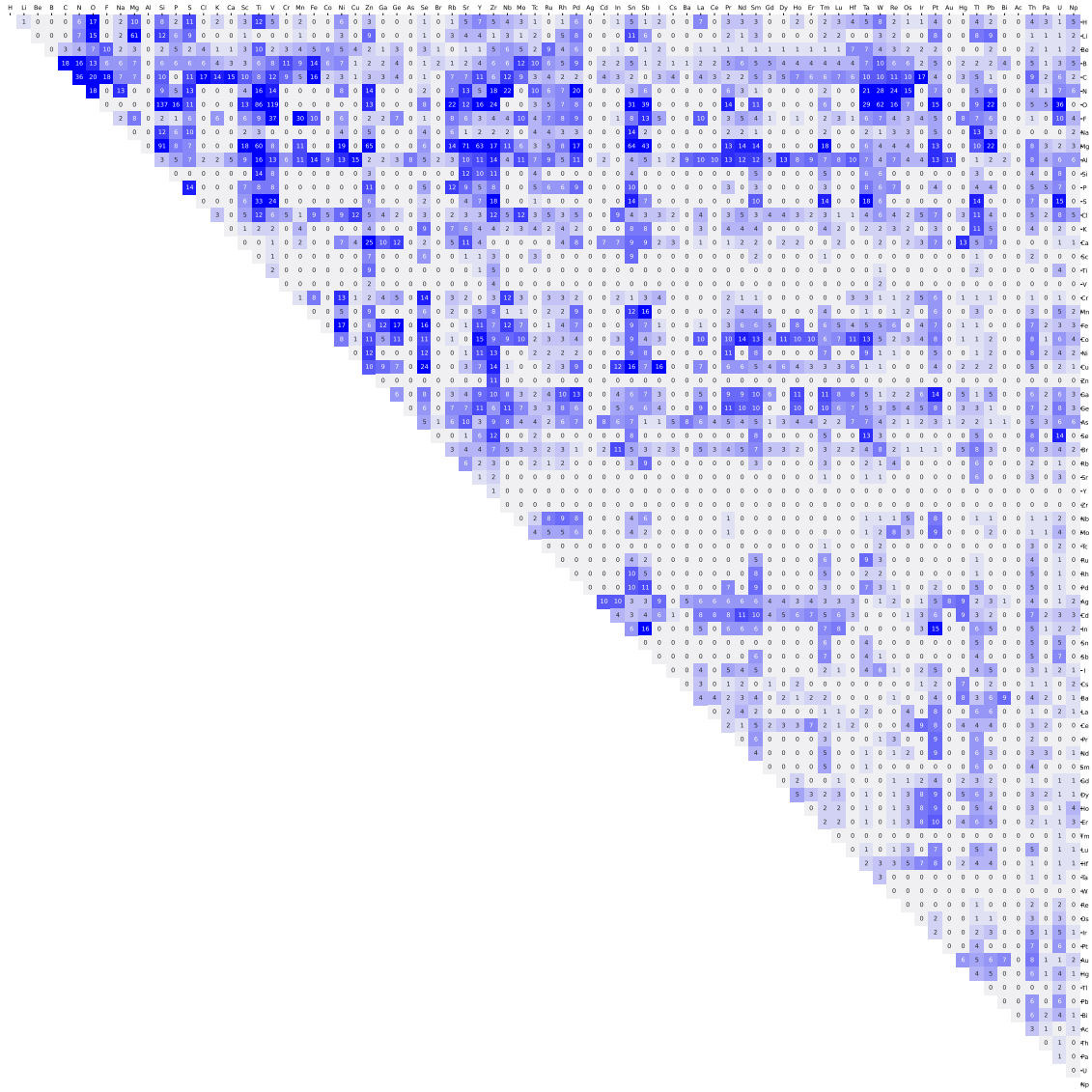
Figura 6.10: Distribuição das energias de gap em função do sistema cristalino para o *dataset* MP-b0. Em (a) não estão inclusos os metais.



Fonte: elaborado pelo autor.

Nas Figuras 6.11 e 6.12 são apresentadas as distribuições de quantidades de compostos para cada par de elemento químico. Nesta representação, o número indica a quantidade de compostos que possuem os elementos químicos determinados pelo cruzamento entre a respectiva linha e respectiva coluna. Essa quantidade reflete compostos com mesma estequiometria, porém sistema cristalino diferente, assim como mesmo sistema cristalino porém proporção entre elementos diferente. Logo a representa apenas endereça a natureza química. Optamos por representar apenas a porção triangular superior da matriz tendo em vista que a ordem de leitura do composto (NaCl ou ClNa) não deveria impactar na quantidade de compostos com aquele par de elementos (Na e Cl). Entretanto, ao observar o *dataset* MP-b0 apresentado na Figura 6.11 nota-se que isso não é garantido já que há várias regiões com lacunas (inclusive para o ClNa que contabiliza zero compostos). Ou seja, a ordem de escrita das fórmulas químicas impacta diretamente na contagem dos compostos, refletindo a preferência humana arbitrária por uma forma de representação.

Figura 6.11: Distribuição dos compostos de acordo com os elementos constituintes para o *dataset* MP-b0.



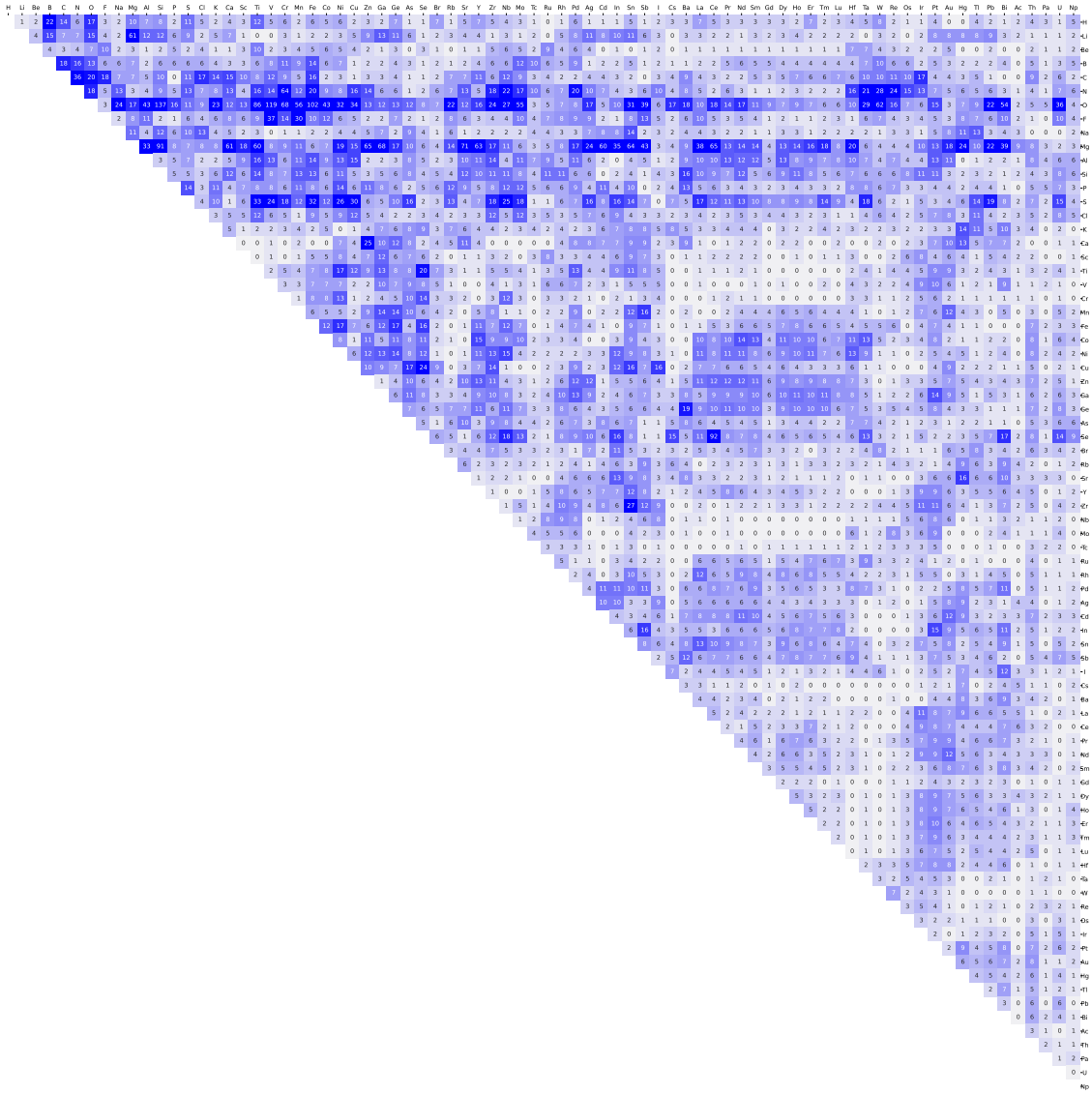
Fonte: elaborado pelo autor.

Já no *dataset* aumentado MP-bA, Figura 6.12, garantimos a presença de todas as representações, dobrando a quantidade de instâncias. Dessa forma verifica-se maior continuidade de distribuição ao longo dos elementos sem que a contagem seja realizada duas vezes para o mesmo composto (já que estamos num corte triangular da matriz - a duplicidade ficaria evidenciada exatamente na região cortada). Agora, as ilhas claras formadas na imagem constituem, de fato, regiões onde há poucos dados, seja porque são compostos pouco estudados na exploração coberta pelo Materials Project, seja porque são compostos que naturalmente não se formam[2] (e portanto foram pouco explorados). A estas regiões correspondem a nomenclatura OOD (*out-of-distribution*). Além disso, nota-se que a medida

[2]Neste caso, nos referimos a energia total positiva.

que avançamos para composições com ambos elementos mais pesados, há uma tendência da intensidade da coloração da Figura 6.12 diminuir, indicando que mesmo para espaços químicos explorados, existem menos dados sobre estruturas (sistemas cristalinos diferentes) e proporções entre elementos.

Figura 6.12: Distribuição dos compostos de acordo com os elementos constituintes para o *dataset* MP-bA.



Fonte: elaborado pelo autor.

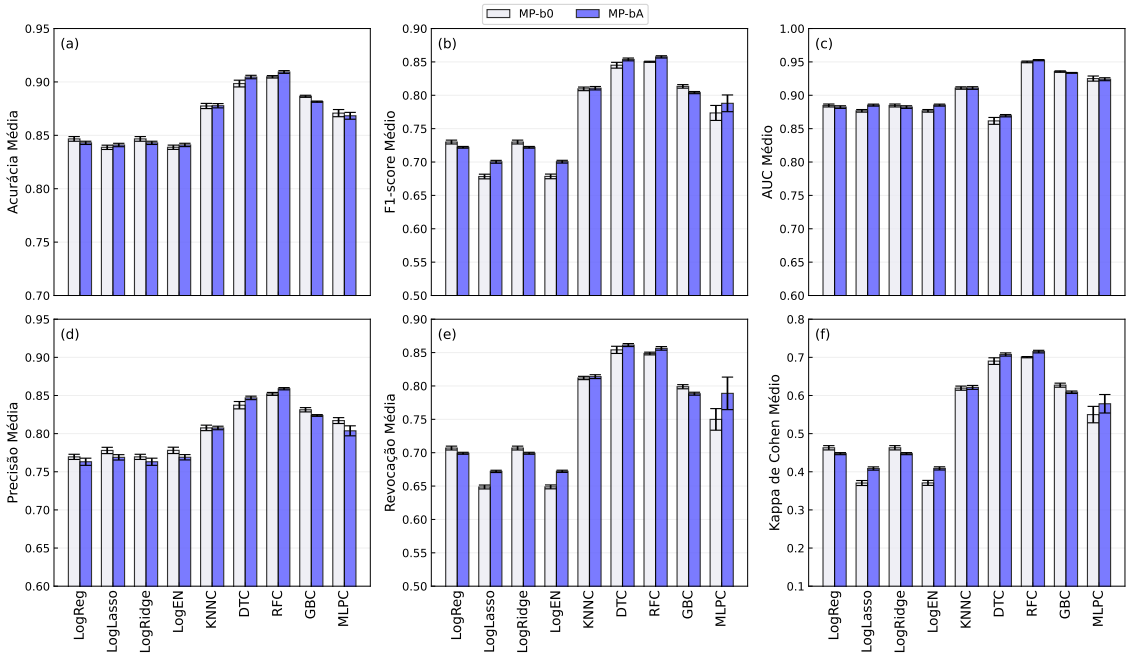
Destacamos que não traz nenhuma informação uma imagem análoga as Figuras 6.11 e 6.12 para o *dataset* GCC-0, uma vez que a contagem em todos os quadrantes será a mesma (devido a geração por combinatória) e igual a 224, resultado da multiplicação entre o número de sistemas cristalinos diferentes (7) e o número de proporções diferentes (32).

6.2.2 Construção dos Modelos de Predição

Os desempenhos observados nos modelos de classificação, tanto para metais e não metais (Figura 6.13), quanto para estáveis e não estáveis (Figura 6.14), demonstram que não há alta variabilidade com relação a validação cruzada sobre todo o *dataset* (barras de incerteza da média pequenas). Isso significa que sob uma escolha aleatória dos dados que irão compor o conjunto de treino e o conjunto de teste, o desempenho não sofre mudanças significativas. Isso ocorre pois numa amostragem aleatória pretende-se manter a estatística original dos dados (amostras de treino e teste devem possuir a mesma distribuição). Também é importante destacar que, para o caso do *dataset* MP-bA, estratificamos as separações para que materiais iguais com representação diferente fossem mantidos no mesmo conjunto, isto é, por exemplo, caso NaCl monoclínico fosse selecionado para o conjunto de treino, ClNa monoclínico também o seria. Com isso pretendíamos não aumentar o desempenho medido por vazar uma informação (do inglês, *data leakage*) entre os conjuntos de treino e teste.

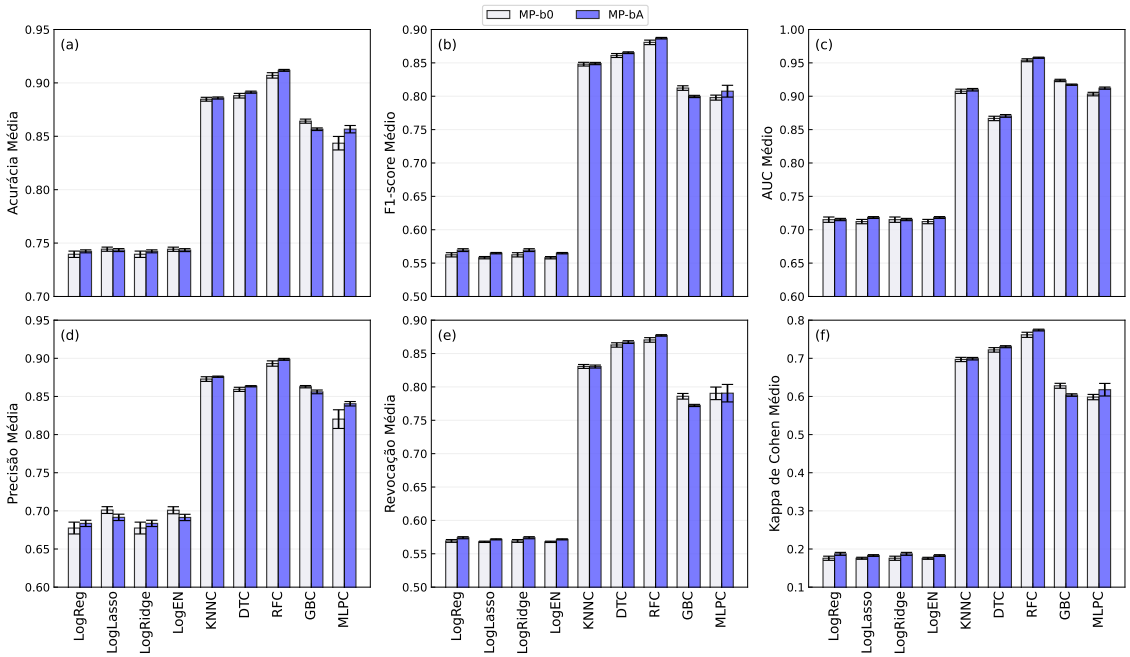
Pela análise desses resultados concluímos que o melhor algoritmo para as tarefas pretendidas é o RandomForest (RFC); o uso do *dataset* aumentado MP-bA é similar ao *dataset* MP-b0, praticamente dentro da incerteza de validação cruzada para a maioria das métricas e algoritmos, com a vantagem de garantir as representações equivalentes; a variabilidade causada pela escolha dos dados de treino e teste é muito pequena, menor que 1% na média.

Figura 6.13: Desempenho dos algoritmos de classificação para determinação de metais e não metais. Precisão, Revocação e F1-score são médias macro sobre as classes. Para todas as métricas, as médias foram calculadas com os resultados de teste de validação cruzada sobre todo o *dataset*. As barras de variação correspondem a incerteza da média.



Fonte: elaborado pelo autor.

Figura 6.14: Desempenho dos algoritmos de classificação para determinação de materiais estáveis ($H_f \leq 0$) e instáveis ($H_f > 0$). Precisão, Revocação e F1-score são médias macro sobre as classes. Para todas as métricas, as médias foram calculadas com os resultados de teste de validação cruzada sobre todo o *dataset*. As barras de variação correspondem a incerteza da média.



Fonte: elaborado pelo autor.

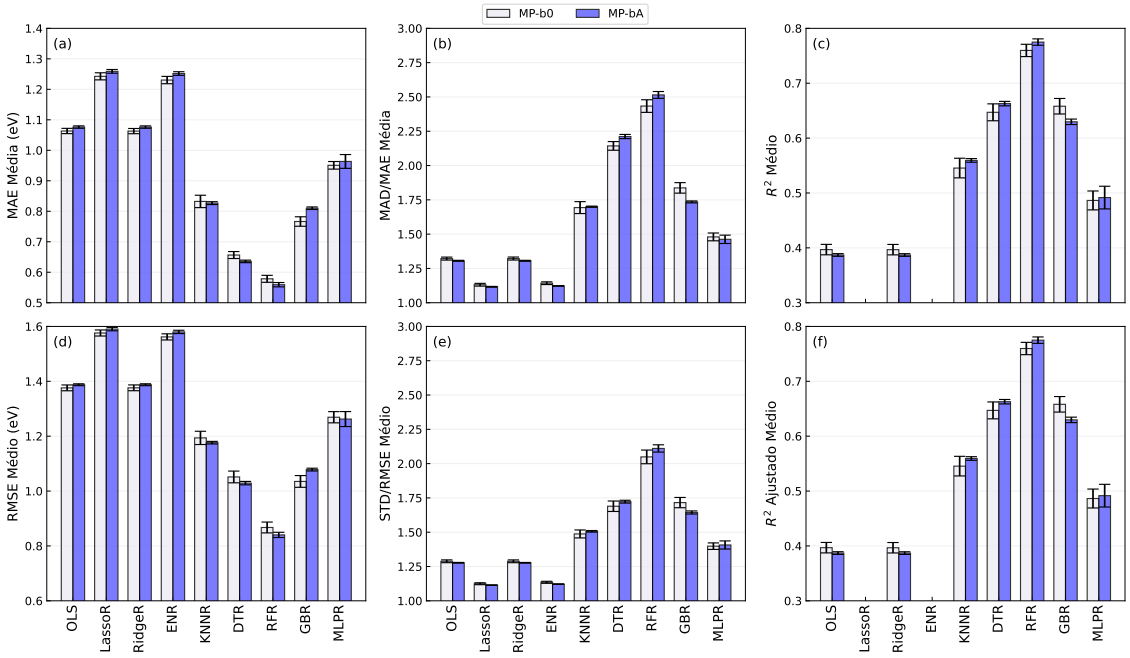
Para estes algoritmos as *features* foram escalonadas de acordo com o algoritmo: para MLPC e KNNC foi usado o `MinMaxScaler`; para DTC, RFC, GBC nenhum escalonador foi usado; para os demais, foi utilizado o `StandardScaler`. O mesmo esquema de escalonamento foi utilizado para os regressores.

Para treinar os regressores utilizamos o mesmo conjunto de *features* dos classificadores. Entretanto, agora usamos apenas dados de semicondutores estáveis ($\Delta H_f \leq 0$) com bandgap maior que 50 meV, totalizando 3042 materiais para MP-b0 e 6084 para MP-bA. Os resultados são apresentados na Figura 6.15.

Constata-se que os modelos lineares tiveram os maiores erros, assim como no caso da classificação. Os melhores modelos foram os baseados em árvores (DTR e RFR). Uma MAD/MAE de 2.5 equivale a dizer que as predições de bandgap foram 2.5 vezes melhores usando o preditor do que seriam ser o bandgap fosse predito pela média dos dados.

Uma MAE média de 0.58 eV significa que o erro médio esperado na predição dos valores de bandgap é dessa ordem. Aqui é importante lembrar que estes modelos foram treinados com bandgap de resultados de simulações DFT PBE-GGA extraídas do Materials Project. Logo, o erro com relação ao bandgap real desses materiais pode ser maior do que a métrica de MAE indica, isto é, devemos levar em consideração que há um deslocamento (subestimação) do bandgap usado no treino com relação ao bandgap real dos materiais.

Figura 6.15: Desempenho dos algoritmos de regressão para predizer o bandgap (E_g) de cada material. Para todas as métricas, as médias foram calculadas com os resultados de teste de validação cruzada sobre todo o *dataset*. As barras de variação correspondem a incerteza da média.



Fonte: elaborado pelo autor.

Todos os modelos aqui discutimos tem espaço para serem melhorados: (1) podemos ajustar os hiperparâmetros para obter uma melhor performance; (2) modificar o conjunto de *features* de entrada, por exemplo, utilizando esquemas de combinação (como soma ou subtração) entre elas, ou de não linearidade (como certas *features* ao quadrado ou ao cubo); (3) podemos modificar as técnicas de escalonamento utilizadas para que capturem certo nuances dos dados (por exemplo, alguns valores são discretos enquanto outros são reais discretizados pelo fato de pertencerem a átomos). Tudo isso indica ricas direções para aperfeiçoamentos futuros.

6.2.3 Interpretabilidade dos Modelos

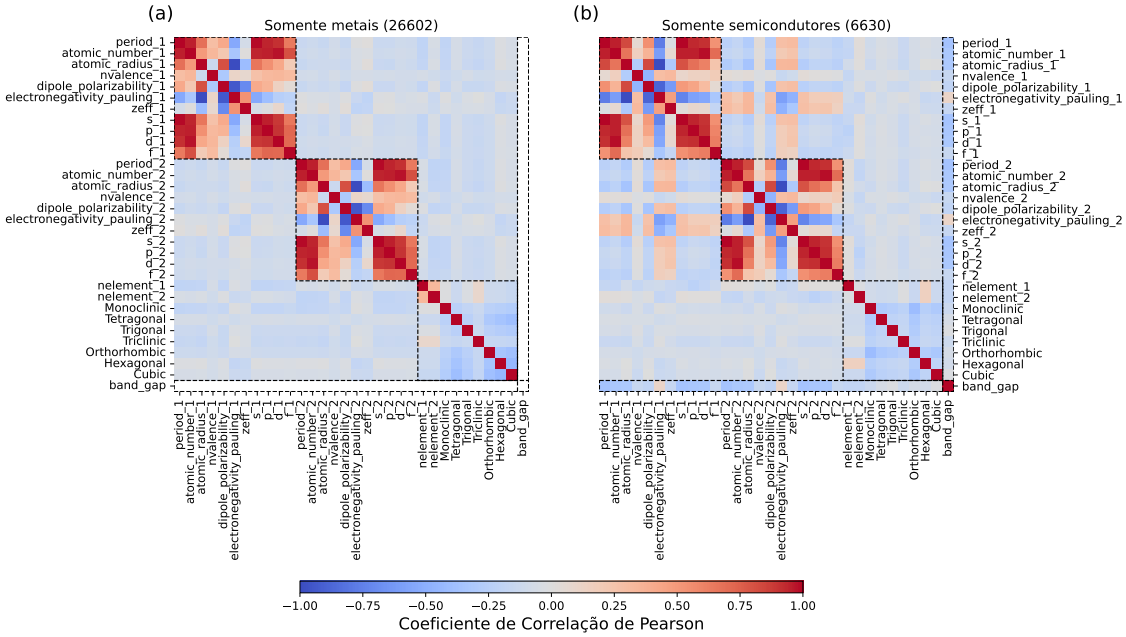
A interpretação e explicabilidade dos modelos finais é de suma importância para que possamos confiar nos resultados gerados. Como discutido na Seção 3.3.2 do Capítulo 3, existem várias técnicas possiveis para realizar isso. Aqui decidimos focar na interpretação da predição dos materiais semicondutores e do bandgap.

O fato de termos usado *features* atômicas auxilia na interpretação física do resultados, uma vez que podemos associar determinados desempenhos a variáveis que são fisicamente estabelecidas.

Na Figura 6.16 é representada a correlação de Pearson entre as *features*. Os quadrantes em tracejado preto representam os agrupamentos de *features* para o primeiro átomo, o

segundo átomo, e aqueles referentes a estrutura. O retângulo tracejado na última linha de cada mapa é a variável de saída bandgap (em branco para os metais já que este valor é nulo para todos os compostos). Foram excluídos semicondutores com bandgap de até 50 meV.

Figura 6.16: Correlação de Pearson entre as *features* para o grupo de metais e de semicondutores.



Fonte: elaborado pelo autor.

A coeficiente de Pearson mede correlação linear entre as variáveis. Neste caso observamos que há alta correlação entre os valores de um mesmo elemento - o que é esperado dado que as propriedades atômicas são únicas. Isto não é válido para os atributos que podem ser o mesmo para mais de um elemento como o número de elétrons de valência (`nvalence`), a polarizabilidade de dipolo (`dipole_polarizability`), a eletronegatividade de Pauling (`ekectronegativity_pauling`) e a carga nuclear efetiva (`zeff`). Estas duas últimas são as que tem menos correlação interna com o próprio elemento, mas apresentam correlação entre si. Observa-se também que a polarizabilidade de dipolo tem relação mais forte com o raio atômico (`atomic_radius`).

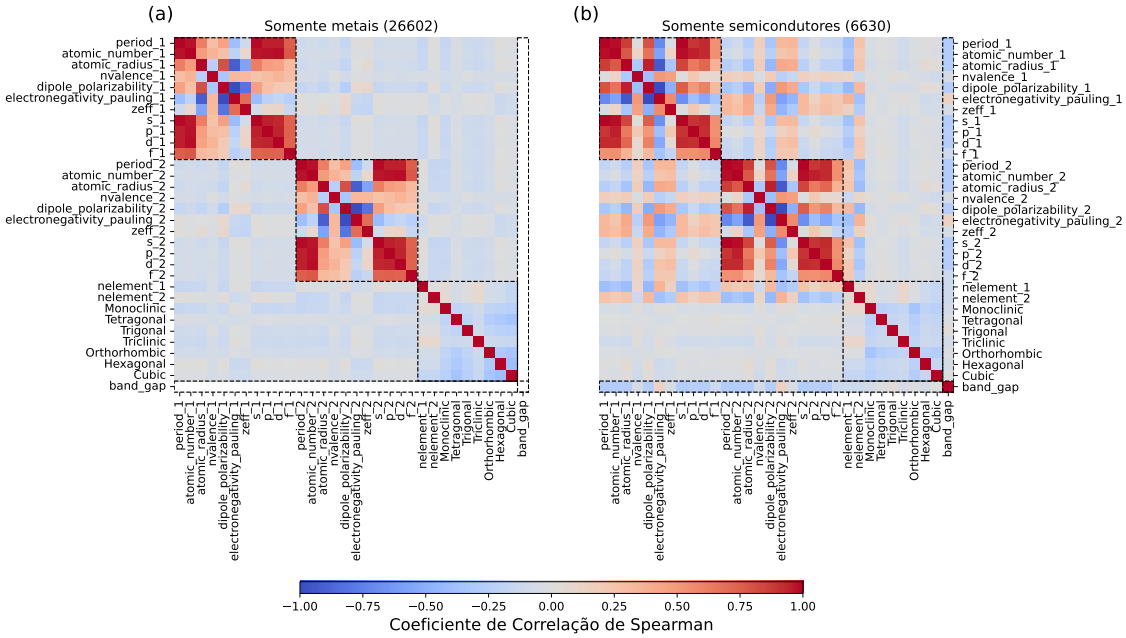
As *features* estruturais quase não apresentam correlação entre si - o que é esperado para o caso do sistema cristalino codificado com OHE - apresentando apenas relação entre o número de átomos de cada espécie quando agrupado por metais.

Mais interessante são as correlações cruzadas entre os dois elementos do composto binário. Para metais, Figura 6.16 (a), ela não aparece; para semicondutores, Figura 6.16 (b), surgem raias no número de valência, eletronegatividade e carga nuclear efetiva. Essas raias representam correlação entre as *features* usadas para representar o primeiro átomo e o segundo átomo. Porém, não são via pares de mesma natureza: valência, eletronegatividade

e carga nuclear efetiva do primeiro átomo se relacionam com período, número atômico, raio atômico, polarizabilidade, e número de elétrons nas camadas s,p,d e f do segundo átomo.

Como a correlação de Pearson só captura relações lineares entre as variáveis, também calculamos a correlação de Spearman, apresentada na Figura 6.17. Esta estatística avalia relações monótonas, sejam elas lineares ou não. Podemos observar que as relações descritas anteriormente se preservam, como era de se esperar. Porém, somente para o semicondutores, surgem raias para o número de átomos de um elemento em cada composto (`nelements`) que agora se relacionam de maneira inversa: a correlação é negativa com as variáveis do átomo trocado e positivas com as variáveis do mesmo átomo.

Figura 6.17: Correlação de Spearman entre as *features* para o grupo de metais e de semicondutores.

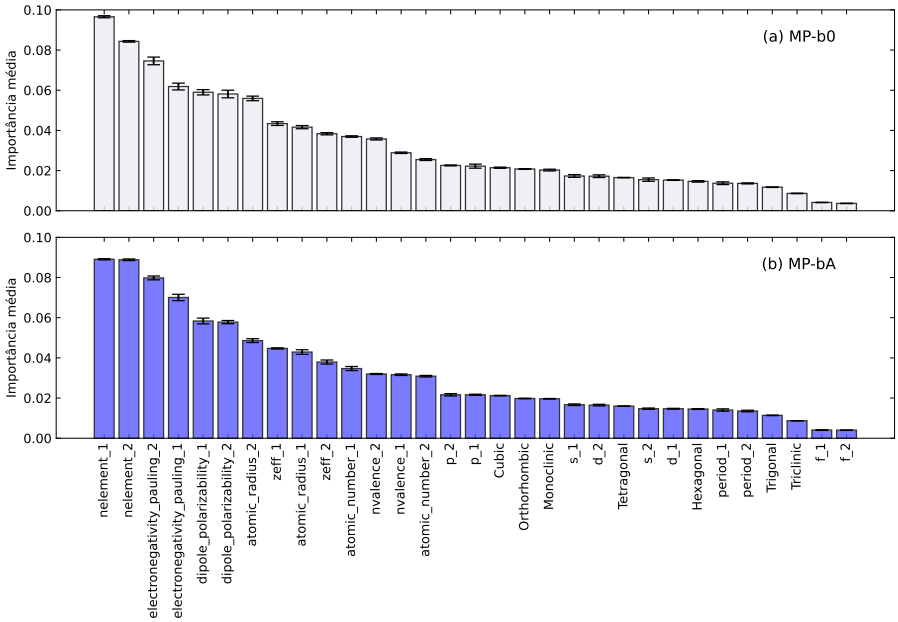


Fonte: elaborado pelo autor.

Para o melhor algoritmo da etapa anterior, no caso a RandomForest, podemos utilizar o atributo `feature_importances`. Este atributo indica, para cada *feature*, a contribuição relativa para reduzir a impureza dos nós na árvore, refletindo a importância de cada variável na previsão do modelo.

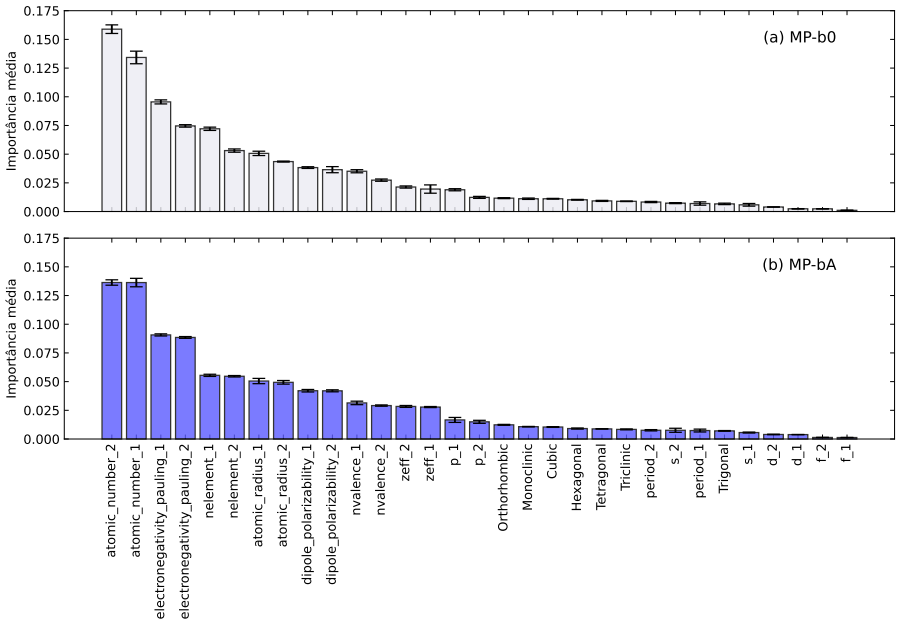
Na Figura 6.18 mostramos esses resultados para o classificador de metais e semicondutores. Comparando o gráfico em (a) e em (b), constatamos que o uso do *dataset* MP-bA equaliza a importância que o algoritmo dá para cada *feature* análoga - i. e., *features* que representam a mesma propriedade física mas para elementos químicos diferentes no composto binário. Além disso, a *feature* `nelement` foi a mais importante, o que está de acordo com as raias observadas na correlação de Spearman, Figura 6.17 (b). O mesmo ocorreu para a segunda *feature* mais importante, `electronegativity_pauling`, e para a terceira `dipole_polarizability`.

Figura 6.18: Importância das *features* para o classificador RFC de metais e semicondutores. Os valores são médias para os 5 testes realizados, com as barras de erro indicando a incerteza da média.



Fonte: elaborado pelo autor.

Figura 6.19: Importância das *features* para o regressor RFR de bandgap. Os valores são médias para os 5 testes realizados, com as barras de erro indicando a incerteza da média.



Fonte: elaborado pelo autor.

Entretanto `zeff`, que possuía raias de contraste bem definidas entre metais e semicon-

dutores nas analises de correlação, foi apenas a quarta *feature*, compartilhando a posição com `atomic_radius`. Acreditamos que isso ocorreu pois `atomic_radius` tem correlação com `dipole_polarizability`, apesar dela em si, não ter sido variável de contraste observada nas correlações.

Na Figura 6.19 repetimos a mesma análise de importância de *features*. Agora é o número atômico a primeira *feature* mais importante, seguido da eletronegatividade e do número de átomos daquele elemento. Esta última é exatamente aquela que se manifesta apenas quando consideramos monotonicidade não-lineares.

6.2.4 Geração de Novos Materiais

Para a geração de novos materiais, utilizamos o *dataset* GCC-0 gerado por combinação química. As entradas foram submetidas aos módulos testadores criados pelos modelos de aprendizado de máquina discutidos acima. Realizamos uma filtragem em cascata: primeiro classificamos o *dataset* inteiro entre estáveis e não estáveis; os materiais estáveis foram classificados entre metais e não metais; por fim, os não metais tiveram o bandgap regredido. O modelo utilizado para cada caso foi a respectiva RandomForest treinada com todos os dados de MP-bA.

Não filtramos os dados de treino contidos em GCC-0 para fazer as predições, mesmo que tais valores pudessem ser mais facilmente aprendido pelo modelo, a quantidade de dados utilizada nessa etapa é muito maior - MP-bA tem até 33550 (sem excluir o limite de bandgap de até 20 meV) enquanto GCC-0 tem 1380298 materiais. De fato estamos explorando uma região muito maior do espaço de materiais, de forma que não esperamos observar um sobreajuste ao analisar todos os dados juntos.

Na Tabela 6.8 são apresentadas as contagens de materiais gerados que foram preditos como estáveis e semicondutores pelos modelos de predição treinados. A maioria são metais estáveis, seguido de metais instáveis, refletindo o desbalanceamento inerente dos dados de treino (Tabela 6.7). Semicondutores gerados totalizam apenas 136853 dos materiais preditos, sendo em sua maioria, estáveis.

Tabela 6.8: Contagem de compostos em GCC-0 separados por tipo de condutividade e por critério de estabilidade termodinâmica.

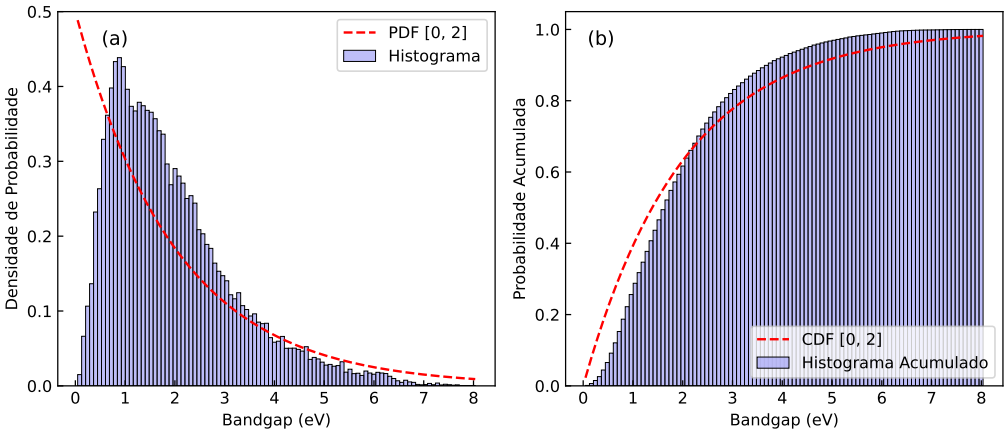
Tipo de Condutividade	Critério de Estabilidade		Total
	$\Delta H_f > 0$	$\Delta H_f \leq 0$	
Metais $E_g = 0$	718516	524919	124761
Semicondutores $E_g > 0$	13647	123206	132667

Fonte: elaborado pelo autor.

Seguindo o sistema em cascata, as predições foram filtradas para selecionar apenas

semicondutores estáveis. Estes materiais foram passados no modelo regressor de bandgap. A Figura 6.20 contém as distribuições de bandgap preditos por esse módulo. Para comparação, traçamos o modelos ajustado sobre os dados. Observa-se que distribuição predita pelo modelo não está igual a distribuição original dos dados. Isso é esperado para modelos de RandomForest pois eles calculam médias de valores na predição, suavizando valores extremos (no caso os valores de bandgap pequeno foram os mais subestimados, i.e., estão longe da curva vermelha dos dados do modelo ajustado nos dados de treino). A curva também demonstra uma concentração no meio, característica de modelos de árvore. Estes são exemplos de limitações inerentes do algoritmo utilizado.

Figura 6.20: Distribuição do bandgap dos semicondutores para o *dataset* GCC-0. (a) Histograma e função densidade de probabilidade (PDF) exponencial ajustada para $E_g > 0$. (b) Histograma acumulado e função de distribuição acumulada (CDF) exponencial ajustada para $E_g > 0$. Para todos os histograma foram usados 99 bins de acordo com a regra de Rice.



Fonte: elaborado pelo autor.

Diversos caminhos são possíveis para entender a diferença entre as curvas aprendidas e generalizadas para a estatística de bandgap. Por exemplo, podemos utilizar informações da Figura 6.12 para selecionar as bolhas de OOD que devem existir em GCC-0, e verificar se eliminando essas bolhas a curva aprendida torna-se mais próxima da original do dados. Ou seja, podemos interpretar como a região de extrapolação do espaço de materiais modifica a distribuição generalizada.

Por fim, mesmo com essas limitações, o nosso sistema pode ser usado na sequência para recomendar estruturas. Podemos selecionar uma determinada composição em GCC-0 que atenda determinado critério, por exemplo, semicondutor estável com bandgap entre 4 eV e 5 eV - isso retorna 5830 composições. Essas composições também tem informação estrutural através do sistema cristalino. Para gerar um arquivo CIF de um candidato, podemos procurar em MP-b0 por estruturas que tenham esse sistema cristalino e a mesma estequiometria da composição sugerida. Se essa estrutura existir, basta substituirmos os respectivos elementos químicos. Em casos de indecisão sobre qual elemento substituir,

podemos utilizar uma heurística de probabilidade de substituição como a escala de Pettifor modificada [134]. Dessa forma temos um material candidato criado por substituição, porém com mais informações acerca da recomendação do que no casos discutido no Capítulo 5.

Esses novos arquivos CIFs podem então passar pelo escrutínio de simulações DFT mais custosas que os métodos de ML usados aqui - vale lembrar que tais métodos não consomem mais do que minutos de processamento em computadores comerciais normais, constituindo uma das vantagens dessa abordagem.

6.3 Conclusão

Neste estudo de caso tivemos a intenção de fazer uma prova de conceito. Demonstramos como é possível utilizar técnicas de aprendizado de máquina para explorar um espaço composicional com informações de sistema cristalino muito mais vasto que nos outros estudos de casos discutidos nessa tese. Logo, este capítulo constitui uma forte expressão do uso da Ciência de Dados de Materiais. Nossas análises se concentraram em questões que deixavam patente o volumes de dados, assim como a necessidade de se usar técnicas estatísticas mais do que técnicas focadas em um único material. O tamanho do espaço de busca que modelos substitutos permitem ter acesso é totalmente diferente daquele que se consegue pela abordagem convencional (Capítulo 4) ou mesmo pelas técnicas de *high throughput* (Capítulo 5). A desvantagem é que abrimos mão da precisão e de um estudo mais aprofundado das propriedades de cada material individual. Devemos, portanto, encarar esses sistemas como guias gerais: mapas em escala maiores, que nos permitem ter ideias de como andar pelo espaço de materiais de maneira, pelo menos, um pouco mais informada do que seria uma simples abordagem por "tentativa e erro".

Outro ponto importante é que mesmo usando técnicas agnósticas à física, pudemos interpretar características físicas: em compostos binários, a semicondutividade parece ter assinaturas atômicas contrastantes com relação aos materiais metálicos. Entre elas temos a relação monotônica entre a estequiometria de cada átomo do composto, a importância da eletronegatividade, a carga nuclear efetiva e a polarizabilidade de dipolo. A correlação entre essas propriedades para os dois átomos em um semicondutor tem um relação distinta com relação a mesma correlação para átomos em um metal.

7 Conclusões

Os estudos de caso deste trabalho foram ordenados de acordo com o grau de mudança de paradigma: lentamente migramos de uma abordagem convencional de experimentação e simulação para uma abordagem orientada por um grande volume de dados. Isso é caracterizado pelo número de materiais em cada etapa: cerca de 5 variações de perovskitas pseudo-2D no Capítulo 4; 100 perovskitas pseudo-1D (jakobssonita) e um pouco menos de 3 mil fases competidoras no Capítulo 5; mais de 16 mil materiais binários (treinamento) e 1 milhão explorados (geração) no Capítulo 6. Essa ordenação também reflete a perspectiva histórica do desenvolvimento da área, com o forte avanço dos repositórios de dados e simulações *high throughput* a partir de 2010 e das técnicas de aprendizado de máquina e inteligência artificial a partir de 2020. É patente, contudo, que uma das principais conclusões que devemos destacar para o design inverso de materiais é que uma abordagem eficiente deve usar essas técnicas exatamente na ordem invertida. As técnicas de IA são úteis pela possibilidade de se explorar uma grande região do espaço de materiais, porém elas não são tão precisas e completas quanto as simulações e também não garantem a realidade física do material - que só pode ser alcançada com a experimentação. Logo, o design inverso deve usar essas técnicas como ponto de partida, selecionando candidatos nunca antes pensados através dos padrões aprendidos por essas ferramentas. Em seguida, devemos aprofundar o estudo dos candidatos, explorando suas propriedades em um maior nível de detalhes: devemos determinar de um ponto de vista teórico a *sintetizabilidade* (propriedades termodinâmicas) e a aplicabilidade (propriedades de interesse) dos candidatos. Munidos com um arcabouço teórico cujas predições são confiáveis, é que devemos realizar a síntese e a caracterização de um novo material.

Não obstante, podemos dividir as principais contribuições deste trabalho em duas vertentes. Quanto ao aprofundamento em uma classe de materiais temos os casos da perovskita híbrida pseudo-2D de Dion-Jacobson e da família da jakobssonita (perovskita pseudo-1D). No primeiro, exploramos em profundidade as propriedades de poucos materiais, porém conseguimos endereçar a correspondência entre resultados teóricos e experimentais. No segundo, não só constatamos as propriedades físicas de novos materiais pertencentes a família de interesse como criamos uma nova base de dados de materiais computacionais. Quanto a questão metodológica, elaboramos uma revisão da área com profunda abrangência e contribuímos com sugestões metodológicas para a criação de sistemas generativos de materiais que retenham capacidade interpretativa e explicativa ao se basearem em recomendações de sistemas puramente preditivos.

Esperamos que tais resultados possam servir para estabelecer a área de Ciência de Dados de Materiais como um novo paradigma indispensável para o estudo e a descoberta de novos materiais.

Os arquivos utilizados nas simulações computacionais e os resultados pré-processados, incluindo os modelos de aprendizado de máquina treinados, estão disponíveis em `https://github.com/ferreirahs/ciencia_dados_materiais`.

Bibliografia

[1] Damian N Valencia. Brief review on COVID-19: The 2020 pandemic caused by SARS-CoV-2. *Cureus*, March 2020.

[2] Organização Pan-Americana da Saúde (OPAS). Oms declara fim da emergência de saúde pública de importância internacional referente à covid-19. `https://www.paho.org/pt/noticias/5-5-2023-oms-declara-fim-da-emergencia-saude-publica-importancia-internacional-referente`, 2023. Acessado em 28 de setembro de 2023.

[3] Edouard Mathieu, Hannah Ritchie, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Joe Hasell, Bobbie Macdonald, Saloni Dattani, Diana Beltekian, Esteban Ortiz-Ospina, and Max Roser. Coronavirus pandemic (covid-19). *Our World in Data*, 2020. https://ourworldindata.org/coronavirus.

[4] Ewen Callaway. The race for coronavirus vaccines: a graphical guide. *Nature*, 580(7805):576–577, April 2020.

[5] Ronaldo Challhua, Larissa Akashi, Jose Zuñiga, Helena Beatriz de Carvalho Ruthner Batista, Ricardo Moratelli, and Ana Champi. Portable reduced graphene oxide biosensor for detection of rabies virus in bats using nasopharyngeal swab samples. *Biosensors and Bioelectronics*, 232:115291, July 2023.

[6] Wladyslaw Kowalski. *Ultraviolet Germicidal Irradiation Handbook*. Springer Berlin Heidelberg, 2009.

[7] Michael Kneissl, Tae-Yeon Seong, Jung Han, and Hiroshi Amano. The emergence and prospects of deep-ultraviolet light-emitting diode technologies. *Nature Photonics*, 13(4):233–244, March 2019.

[8] Milad Raeiszadeh and Babak Adeli. A critical review on ultraviolet disinfection systems against COVID-19 outbreak: Applicability, validation, and safety considerations. *ACS Photonics*, 7(11):2941–2951, October 2020.

[9] M. A. S. Pessôa, F. M. e Silva, M. P. Lima Jr., G. Galhardo, P. H. M. Olyntho, and A. A. R. Neves. A física de um desinfector com radiação UV–c. *Revista Brasileira de Ensino de Física*, 43, 2021.

[10] M . Asif Khan; Qhalid Fareed; Vinod Adivarahan. Ultraviolet light emitting devices and methods of fabrication, U.S. Patent 9,985,177 B2, May. 2018.

[11] Masato Toita; Jianfeng Chen; Yuxin Li; Yuting Wang; Hironori Ishii; Ken Kitamura. Ultraviolet light-emitting devices and methods, U.S. Patent 10,383,963 B2, Aug. 2019.

[12] Ray-Hua Horng, Yan-Chung Lai, and Liang-Hsing Lai. Deep-ultraviolet LEDs fabricated by nanoimprinting. *ECS Journal of Solid State Science and Technology*, 9(1):015005, October 2019.

[13] Safa O Kasap. *Optoelectronics and photonics: Principles and practices*. Pearson, 2001.

[14] Mark Fox. *Optical properties of solids*. Oxford Master Series in Physics. Oxford University Press, London, England, 2 edition, March 2010.

[15] Joo Sung Kim, Jung-Min Heo, Gyeong-Su Park, Seung-Je Woo, Changsoon Cho, Hyung Joong Yun, Dong-Hyeok Kim, Jinwoo Park, Seung-Chul Lee, Sang-Hwan Park, Eojin Yoon, Neil C. Greenham, and Tae-Woo Lee. Ultra-bright, efficient and stable perovskite light-emitting diodes. *Nature*, 611(7937):688–694, November 2022.

[16] Yuanzhi Jiang, Junli Wei, and Mingjian Yuan. Energy-funneling process in quasi-2d perovskite light-emitting diodes. *The Journal of Physical Chemistry Letters*, 12(10):2593–2606, March 2021.

[17] Yitong Dong, Ya-Kun Wang, Fanglong Yuan, Andrew Johnston, Yuan Liu, Dongxin Ma, Min-Jae Choi, Bin Chen, Mahshid Chekini, Se-Woong Baek, Laxmi Kishore Sagar, James Fan, Yi Hou, Mingjian Wu, Seungjin Lee, Bin Sun, Sjoerd Hoogland, Rafael Quintero-Bermudez, Hinako Ebe, Petar Todorovic, Filip Dinic, Peicheng Li, Hao Ting Kung, Makhsud I. Saidaminov, Eugenia Kumacheva, Erdmann Spiecker, Liang-Sheng Liao, Oleksandr Voznyy, Zheng-Hong Lu, and Edward H. Sargent. Bipolar-shell resurfacing for blue LEDs based on strongly confined perovskite quantum dots. *Nature Nanotechnology*, 15(8):668–674, July 2020.

[18] Chen Ziming and et al. Roadmap on perovskite light-emitting diodes, November 2023.

[19] Gabriel R Schleder, Antonio C M Padilha, Carlos Mera Acosta, Marcio Costa, and Adalberto Fazzio. From DFT to machine learning: recent approaches to materials science–a review. *Journal of Physics: Materials*, 2(3):032001, May 2019.

[20] Eric S. Lander et. al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, February 2001.

[21] U.S. Federal government. Materials genome initiative. `https://www.mgi.gov/`, 2021. Acessado em 6 de novembro de 2023.

[22] Steven S Skiena. *The Data Science Design Manual*. Texts in computer science. Springer International Publishing, Cham, Switzerland, 1 edition, July 2017.

[23] Isao Tanaka, Krishna Rajan, and Christopher Wolverton. Data-centric science for materials innovation. *MRS Bulletin*, 43(9):659–663, September 2018.

[24] Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, July 2018.

[25] Luigi Sbailò, Ádám Fekete, Luca M. Ghiringhelli, and Matthias Scheffler. The NOMAD artificial-intelligence toolkit: turning materials-science data into knowledge and understanding. *npj Computational Materials*, 8(1), December 2022.

[26] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4), December 2019.

[27] Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3), July 2022.

[28] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, July 2021.

[29] Zeljko Ivezic, Andrew J Connolly, Jacob VanderPlas, and Alexander Gray. *Statistics, data mining, and machine learning in astronomy.* Princeton Series in Modern Observational Astronomy. Princeton University Press, Princeton, NJ, January 2014.

[30] R.K. Bock, A. Chilingarian, M. Gaug, F. Hakl, T. Hengstebeck, M. Jiřina, J. Klaschka, E. Kotrč, P. Savický, S. Towers, A. Vaiciulis, and W. Wittek. Methods for multidimensional event classification: a case study using images from a cherenkov gamma-ray telescope. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 516(2-3):511–528, January 2004.

[31] Sean R Eddy. " antedisciplinary " science. *PLoS Computational Biology*, 1(1):e6, 2005.

[32] Gabriel R. Schleder and Adalberto Fazzio. Machine learning na física, química, e ciência de materiais: Descoberta e design de materiais. *Revista Brasileira de Ensino de Física*, 43(suppl 1), 2021.

[33] Norbert Wiener. *The human use of human beings.* Da Capo Press, London, England, March 1988.

[34] Charles Kittel. *Introduction to solid state physics.* John Wiley & Sons, Nashville, TN, 8 edition, October 2004.

[35] Yongliang Wang. *Basic Theory of Finite Element Method.* EDP Sciences, 2023.

[36] Sinan Muftu. *Finite element method.* Academic Press, San Diego, CA, July 2022.

[37] Nelson H Morgon. *Métodos de Química Teórica E Modelagem Molecular.* Editora Livraria da Fisica, 2007.

[38] David Jeffrey Griffiths. *Mecânica quântica.* Editora Pearson, 2011.

[39] M. Born and R. Oppenheimer. Zur quantentheorie der molekeln. *Annalen der Physik*, 389(20):457–484, 1927.

[40] Nicola Marzari, Andrea Ferretti, and Chris Wolverton. Electronic-structure methods for materials design. *Nature Materials*, 20(6):736–749, May 2021.

[41] R. P. Feynman. Forces in molecules. *Phys. Rev.*, 56:340–343, Aug 1939.

[42] Jose David M Vianna, Adalberto Fazzio, and Sylvio Canuto. *Teoria quântica de moléculas e sólidos: simulação computacional.* Editora Livraria da Física, 2004.

[43] Radu Iftimie, Peter Minary, and Mark E. Tuckerman. Ab initio molecular dynamics: Concepts, recent developments, and future trends. *Proceedings of the National Academy of Sciences*, 102(19):6654–6659, 2005.

[44] User Editor, Antônio S. N. Aguiar, Ademir J. Camargo, and Solemar S. Oliveira. Dinâmica molecular de car-parrinello. *Revista Processos Químicos*, 3(5):59–72, January 2009.

[45] Miguel Angelo F. de Souza and Ricardo L. Longo. Dinâmica molecular de born-oppenheimer: metodologia e aplicações em mecanismos e seletividades de reações químicas. In *Aplicações de química teórica no estudo de materiais: métodos in silico para nanomateriais*, pages 14–61. EdUFSCar, 2018.

[46] J. C. Slater. The theory of complex spectra. *Physical Review*, 34(10):1293–1322, November 1929.

[47] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev.*, 136:B864–B871, Nov 1964.

[48] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev.*, 140:A1133–A1138, Nov 1965.

[49] Paul Adrien Maurice Dirac and Ralph Howard Fowler. Quantum mechanics of many-electron systems. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 123(792):714–733, April 1929.

[50] Richard M Martin. Preface. In *Electronic Structure*, pages xvii–xix. Cambridge University Press, Cambridge, April 2004.

[51] Pedro Borlido, Jonathan Schmidt, Ahmad W. Huran, Fabien Tran, Miguel A. L. Marques, and Silvana Botti. Exchange-correlation functionals for band gaps of solids: benchmark, reparametrization and machine learning. *npj Computational Materials*, 6(1), July 2020.

[52] Robert G. Parr and Yang Weitao. *Density-Functional Theory of Atoms and Molecules*. Oxford University Press, January 1995.

[53] G. L. Oliver and J. P. Perdew. Spin-density gradient expansion for the kinetic energy. *Phys. Rev. A*, 20:397–403, Aug 1979.

[54] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77:3865–3868, Oct 1996.

[55] John P. Perdew, Adrienn Ruzsinszky, Gábor I. Csonka, Oleg A. Vydrov, Gustavo E. Scuseria, Lucian A. Constantin, Xiaolan Zhou, and Kieron Burke. Restoring the density-gradient expansion for exchange in solids and surfaces. *Physical Review Letters*, 100(13), April 2008.

[56] James W. Furness, Aaron D. Kaplan, Jinliang Ning, John P. Perdew, and Jianwei Sun. Accurate and numerically efficient r2scan meta-generalized gradient approximation. *The Journal of Physical Chemistry Letters*, 11(19):8208–8215, September 2020.

[57] Ryan Kingsbury, Ayush S. Gupta, Christopher J. Bartel, Jason M. Munro, Shyam Dwaraknath, Matthew Horton, and Kristin A. Persson. Performance comparison of r2scan and scan metagga density functionals for solid materials via an automated, high-throughput computational workflow. *Physical Review Materials*, 6(1), January 2022.

[58] Jochen Heyd, Gustavo E. Scuseria, and Matthias Ernzerhof. Hybrid functionals based on a screened coulomb potential. *The Journal of Chemical Physics*, 118(18):8207–8215, 2003.

[59] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A*, 38(6):3098–3100, September 1988.

[60] Chengteh Lee, Weitao Yang, and Robert G. Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2):785–789, January 1988.

[61] Lucia Reining. The gw approximation: content, successes and limitations. *WIREs Computational Molecular Science*, 8(3), December 2017.

[62] Samuel Mendes de Vasconcelos. Aproximação gw: Fundamentos e aplicações a novas nanoestruturas bidimensionais. `http://hdl.handle.net/1843/43183`. Acessado em 14 de abril de 2025.

[63] Mariana M. Odashima, Beatriz G. Prado, and E. Vernek. Pedagogical introduction to equilibrium green's functions: condensed-matter examples with numerical implementations. *Revista Brasileira de Ensino de Física*, 39(1), September 2016.

[64] Erich Runge and E. K. U. Gross. Density-functional theory for time-dependent systems. *Physical Review Letters*, 52(12):997–1000, March 1984.

[65] J. Schirmer. Review of the foundations of time-dependent density-functional theory (tddft). *Physical Chemistry Chemical Physics*, 27(10):4992–5005, 2025.

[66] T. J. Zuehlsdorff, N. D. M. Hine, J. S. Spencer, N. M. Harrison, D. J. Riley, and P. D. Haynes. Linear-scaling time-dependent density-functional theory in the linear response formalism. *The Journal of Chemical Physics*, 139(6), August 2013.

[67] Thiago Puccinelli, Murilo Sodré Marques, and José Rafael Bordin. O pacote lammps como uma ferramenta para o ensino de transição de fases. *Revista Brasileira de Ensino de Física*, 46, 2024.

[68] Aidan P. Thompson, H. Metin Aktulga, Richard Berger, Dan S. Bolintineanu, W. Michael Brown, Paul S. Crozier, Pieter J. in 't Veld, Axel Kohlmeyer, Stan G. Moore, Trung Dac Nguyen, Ray Shan, Mark J. Stevens, Julien Tranchida, Christian Trott, and Steven J. Plimpton. Lammps - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications*, 271:108171, February 2022.

[69] Loup Verlet. Computer "experiments" on classical fluids. i. thermodynamical properties of lennard-jones molecules. *Physical Review*, 159(1):98–103, July 1967.

[70] Judith A. Harrison, J. David Schall, Sabina Maskey, Paul T. Mikulski, M. Todd Knippenberg, and Brian H. Morrow. Review of force fields and intermolecular potentials used in atomistic computational materials research. *Applied Physics Reviews*, 5(3), August 2018.

[71] Sören von Bülow, Jakob Tómas Bullerjahn, and Gerhard Hummer. Systematic errors in diffusion coefficients from long-time molecular dynamics simulations at constant pressure. *The Journal of Chemical Physics*, 153(2), July 2020.

[72] Lucas Martin Farigliano, Fabio Negreiros Ribeiro, and Gustavo Martini Dalpian. Phase transitions in cspbbr3: evaluating perovskite behavior over different time scales. *Materials Advances*, 5(14):5794–5801, 2024.

[73] Josep Gelpi, Adam Hospital, Ramón Goñi, and Modesto Orozco. Molecular dynamics simulations: advances and applications. *Advances and Applications in Bioinformatics and Chemistry*, page 37, November 2015.

[74] Sebastian Kmiecik, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski. Coarse-grained protein models and their applications. *Chemical Reviews*, 116(14):7898–7936, June 2016.

[75] J. W. Gibbs. On the equilibrium of heterogeneous substances. *American Journal of Science*, s3-16(96):441–458, 1878.

[76] Robert DeHoff. *Thermodynamics in Materials Science*. CRC Press, Boca Raton, FL, 2 edition, March 2006.

[77] Sicong Jiang and Kesong Yang. Review of high-throughput computational design of heusler alloys. *Journal of Alloys and Compounds*, 867:158854, June 2021.

[78] Artem R. Oganov, Chris J. Pickard, Qiang Zhu, and Richard J. Needs. Structure prediction drives materials discovery. *Nature Reviews Materials*, 4(5):331–348, April 2019.

[79] Wenhao Sun, Stephen T. Dacek, Shyue Ping Ong, Geoffroy Hautier, Anubhav Jain, William D. Richards, Anthony C. Gamst, Kristin A. Persson, and Gerbrand Ceder. The thermodynamic scale of inorganic crystalline metastability. *Science Advances*, 2(11), November 2016.

[80] Vladan Stevanović, Stephan Lany, Xiuwen Zhang, and Alex Zunger. Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies. *Physical Review B*, 85(11), March 2012.

[81] Rico Friedrich, Demet Usanmaz, Corey Oses, Andrew Supka, Marco Fornari, Marco Buongiorno Nardelli, Cormac Toher, and Stefano Curtarolo. Coordination corrected ab initio formation enthalpies. *npj Computational Materials*, 5(1), May 2019.

[82] Muratahan Aykol, Shyam S. Dwaraknath, Wenhao Sun, and Kristin A. Persson. Thermodynamic limit for synthesis of metastable inorganic materials. *Science Advances*, 4(4), April 2018.

[83] Ioanna Pallikara, Prakriti Kayastha, Jonathan M Skelton, and Lucy D Whalley. The physical significance of imaginary phonon modes in crystals. *Electronic Structure*, 4(3):033002, July 2022.

[84] R.P. Futrelle and D.J. McGinty. Calculation of spectra and correlation functions from molecular dynamics data using the fast fourier transform. *Chemical Physics Letters*, 12(2):285–287, December 1971.

[85] Atsushi Togo and Isao Tanaka. First principles phonon calculations in materials science. *Scripta Materialia*, 108:1–5, November 2015.

[86] Martin Thomas, Martin Brehm, Reinhold Fligg, Peter Vöhringer, and Barbara Kirchner. Computing vibrational spectra from ab initio molecular dynamics. *Physical Chemistry Chemical Physics*, 15(18):6608, 2013.

[87] Yonggang Zhao, Kunhao Li, and Jing Li. Review. solvothermal synthesis of multifunctional coordination polymers. *Zeitschrift für Naturforschung B*, 65(8):976–998, August 2010.

[88] Huayang Li, Nan Shen, Shi Chen, Fei Guo, and Baomin Xu. Recent progress on synthesis, intrinsic properties and optoelectronic applications of perovskite single crystals. *Advanced Functional Materials*, 33(24), March 2023.

[89] C. Jeffrey Brinker and George W. Scherer. *Sol-Gel Science: The Physics and Chemistry of Sol–Gel Processing*. Elsevier, 1990.

[90] Larry L. Hench and Jon K. West. The sol-gel process. *Chemical Reviews*, 90(1):33–72, January 1990.

[91] John P. Ciraldo, Jonathan Levine, and Hasitha Ganegoda. Improvements to crystal quality of sapphire grown by the kyropoulos method. In Sergei G. Romanov, Gabriel Lozano, Dario Gerace, Christelle Monat, and Hernán R. Míguez, editors, *Photonic Crystal Materials and Devices XI*, volume 9127, page 91271C. SPIE, May 2014.

[92] Ibrahim Elswie, Zorica Lazarevic, Vesna Radojevic, Martina Gilic, Maja Rabasovic, Dragutin Sevic, and Nebojsa Romcevic. The bridgman method growth and spectroscopic characterization of calcium fluoride single crystals. *Science of Sintering*, 48(3):333–341, 2016.

[93] Zbigniew Galazka, Klaus Irmscher, Reinhard Uecker, Rainer Bertram, Mike Pietsch, Albert Kwasniewski, Martin Naumann, Tobias Schulz, Robert Schewski, Detlef Klimm, and Matthias Bickermann. On the bulk $\beta-Ga_2O_3$ single crystals grown by the czochralski method. *Journal of Crystal Growth*, 404:184–191, October 2014.

[94] Tetsuya Yamada, Hiromasa Kaneko, Fumitaka Hayashi, Tatsuya Doi, Michihisa Koyama, and Katsuya Teshima. Development of a flux-method process informatics system and its application in growth control for layered perovskite ba5nb4o15 crystals. *Crystal Growth and Design*, 23(12):8678–8693, October 2023.

[95] P. S. Vadilo. *Purification of Metals by Zone Melting*, page 197–199. Springer New York, 1975.

[96] Zhengyang Cai, Bilu Liu, Xiaolong Zou, and Hui-Ming Cheng. Chemical vapor deposition growth and applications of two-dimensional materials and their heterostructures. *Chemical Reviews*, 118(13):6091–6133, January 2018.

[97] Massao Ionashiro and I Giolito. Fundamentos da termogravimetria, análise térmica diferencial, calorimetria exploratória diferencial. *São Paulo: Giz Editorial*, 2005.

[98] Elton N Kaufmann, editor. *Characterization of materials, 2 volume set.* John Wiley & Sons, Nashville, TN, March 2003.

[99] Mosayeb Naseri, Sergey Gusarov, and D. R. Salahub. Quantum machine learning in materials prediction: A case study on abo3 perovskite structures. *The Journal of Physical Chemistry Letters*, 14(31):6940–6947, July 2023.

[100] Rosangela Marquesone. *Big Data - Técnicas e tecnologias para extração de valor dos dados.* Casa do Código, 2016.

[101] Robert W. Epps, Michael S. Bowen, Amanda A. Volk, Kameel Abdel-Latif, Suyong Han, Kristofer G. Reyes, Aram Amassian, and Milad Abolhasani. Artificial chemist: An autonomous quantum dot synthesis bot. *Advanced Materials*, 32(30), June 2020.

[102] B. P. MacLeod, F. G. L. Parlane, T. D. Morrissey, F. Häse, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. E. Yunker, M. B. Rooney, J. R. Deeth, V. Lai, G. J. Ng, H. Situ, R. H. Zhang, M. S. Elliott, T. H. Haley, D. J. Dvorak, A. Aspuru-Guzik, J. E. Hein, and C. P. Berlinguette. Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances*, 6(20), May 2020.

[103] Patrick Henry Winston. *Inteligência Artificial.* LTC, 1987.

[104] Elaine Rich. *Inteligência Artificial.* McGraw Hill, 1988.

[105] Diane E. Drang e Barry Edelson Robert I. Levine. *Inteligência Artificial e Sistemas Especialistas.* McGraw Hill, 1988.

[106] João de Fernandes Teixeira. *Inteligência Artificial.* Paulus Editora, 2009.

[107] Ben Coppin. *Inteligência Artificial.* LTC - GRUPO GEN, 2010.

[108] A. M. TURING. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236):433–460, October 1950.

[109] Stuart Russel e Peter Norvig. *Inteligência Artificial.* GEN LTC, 2013.

[110] Charbel Niño El-Hani and Antonio Augusto Passos Videira. *O que é vida?: para entender a biologia do século XXI.* Editora Faperj, 2000.

[111] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to Algorithms.* The MIT Press. MIT Press, London, England, 3 edition, July 2009.

[112] Kyle Hamilton, Aparna Nayak, Bojan Božić, and Luca Longo. Is neuro-symbolic AI meeting its promises in natural language processing? a structured review. *Semantic Web*, pages 1–42, November 2022.

[113] Joel S Cohen. *Computer algebra and symbolic computation.* CRC Press, Boca Raton, FL, July 2002.

[114] Robert K. Lindsay, Bruce G. Buchanan, Edward A. Feigenbaum, and Joshua Lederberg. DENDRAL: A case study of the first expert system for scientific hypothesis formation. *Artificial Intelligence*, 61(2):209–261, June 1993.

[115] Michael Sipser. *Introdução à teoria da computação.* Cengage Learning, 2007.

[116] Stephen Cook. The p versus np problem. `https://www.claymath.org/wp-content/uploads/2022/06/pvsnp.pdf`, 2022. Acessado em 24 de novembro de 2023.

[117] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229, July 1959.

[118] André Carlos Ponce de Leon Ferreira et al. CARVALHO. *Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina.* LTC, March 2021.

[119] Damjan Krstajic, Ljubomir J Buturovic, David E Leahy, and Simon Thomas. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, 6(1), March 2014.

[120] Andriy Burkov. *Machine learning engineering.* True Positive, September 2020.

[121] Suorong Yang, Weikang Xiao, Mengchen Zhang, Suhan Guo, Jian Zhao, and Furao Shen. Image data augmentation for deep learning: A survey, 2022.

[122] Guilherme Aldeia, Henrique Ferreira, and Kenji Filho. On the semblance based tdoa algorithm for sound source localization: a parametric study. In *Anais de XXXVIII Simpósio Brasileiro de Telecomunicações e Processamento de Sinais*, SBrT2020. Sociedade Brasileira de Telecomunicações, 2020.

[123] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, June 2002.

[124] Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1), November 2021.

[125] Nambury S. Raju, Reyhan Bilgic, Jack E. Edwards, and Paul F. Fleer. Methodology review: Estimation of population validity and cross-validity, and the use of equal weights in prediction. *Applied Psychological Measurement*, 21(4):291–305, December 1997.

[126] Jan Brabec, Tomáš Komárek, Vojtěch Franc, and Lukáš Machlica. *On Model Evaluation Under Non-constant Class Imbalance*, page 74–87. Springer International Publishing, 2020.

[127] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, April 1960.

[128] Nuwan Perera and Rangika S. Hikkaduwa Koralege. *High throughput screening*, page 297–301. Elsevier, 2024.

[129] Highlights from the E-CAM Project ; E-CAM — e-cam2020.eu. `https://www.e-cam2020.eu/highlights-from-e-cam/`. Acessado em 24 de novembro de 2023.

[130] X. D. Xiang, Xiaodong Sun, Gabriel Briceño, Yulin Lou, Kai-An Wang, Hauyee Chang, William G. Wallace-Freedman, Sung-Wei Chen, and Peter G. Schultz. A combinatorial approach to materials discovery. *Science*, 268(5218):1738–1740, June 1995.

[131] Wahyu Setyawan and Stefano Curtarolo. High-throughput electronic band structure calculations: Challenges and tools. *Computational Materials Science*, 49(2):299–312, August 2010.

[132] Xiangyang Liu, Haiyang Niu, and Artem R. Oganov. Copex: co-evolutionary crystal structure prediction algorithm for complex systems. *npj Computational Materials*, 7(1), December 2021.

[133] Hai-Chen Wang, Silvana Botti, and Miguel A. L. Marques. Predicting stable crystalline compounds using chemical similarity. *npj Computational Materials*, 7(1), January 2021.

[134] Henning Glawe, Antonio Sanna, E K U Gross, and Miguel A L Marques. The optimal one dimensional periodic table: a modified pettifor chemical scale from data mining. *New Journal of Physics*, 18(9):093011, September 2016.

[135] Ryan S. Kingsbury, Andrew S. Rosen, Ayush S. Gupta, Jason M. Munro, Shyue Ping Ong, Anubhav Jain, Shyam Dwaraknath, Matthew K. Horton, and Kristin A. Persson. A flexible and scalable scheme for mixing computed formation energies from different levels of theory. *npj Computational Materials*, 8(1), September 2022.

[136] Koichi Momma and Fujio Izumi. Vesta: a three-dimensional visualization system for electronic and structural analysis. *Journal of Applied Crystallography*, 41(3):653–658, May 2008.

[137] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source

python library for materials analysis. *Computational Materials Science*, 68:314–319, February 2013.

[138] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, Eric D Hermes, Paul C Jennings, Peter Bjerre Jensen, James Kermode, John R Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W Jacobsen. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, June 2017.

[139] Vei Wang, Nan Xu, Jin-Cheng Liu, Gang Tang, and Wen-Tong Geng. Vaspkit: A user-friendly interface facilitating high-throughput computing and analysis using vasp code. *Computer Physics Communications*, 267:108033, 2021.

[140] Giovanni Pizzi, Andrea Cepellotti, Riccardo Sabatini, Nicola Marzari, and Boris Kozinsky. Aiida: automated interactive infrastructure and database for computational science. *Computational Materials Science*, 111:218–230, January 2016.

[141] William Gropp, Ewing Lusk, and Anthony Skjellum. *Using MPI*. Scientific and Engineering Computation. MIT Press, London, England, 2 edition, November 1999.

[142] Andy B. Yoo, Morris A. Jette, and Mark Grondona. *SLURM: Simple Linux Utility for Resource Management*, page 44–60. Springer Berlin Heidelberg, 2003.

[143] Douglas Thain, Todd Tannenbaum, and Miron Livny. Distributed computing in practice: the condor experience: Research articles. *Concurr. Comput.: Pract. Exper.*, 17(2–4):323–356, February 2005.

[144] Tom White. *Hadoop - The Definitive Guide 4e*. O'Reilly Media, Sebastopol, CA, May 2015.

[145] Bill Chambers and Matei Zaharia. *Spark - the definitive guide*. O'Reilly Media, Sebastopol, CA, March 2018.

[146] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1), July 2013.

[147] Scott Kirklin, James E Saal, Bryce Meredig, Alex Thompson, Jeff W Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The open quantum materials

database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1(1), December 2015.

[148] Mariette Hellenbrandt. The inorganic crystal structure database (icsd)—present and future. *Crystallography Reviews*, 10(1):17–22, January 2004.

[149] Emilio Vital Brazil, Eduardo Soares, Lucas Villa Real, Leonardo Azevedo, Vinicius Segura, Luiz Zerkowski, and Renato Cerqueira. Position paper on dataset engineering to accelerate science, 2023.

[150] Tim Berners-Lee and James Hendler. Publishing on the semantic web. *Nature*, 410(6832):1023–1024, April 2001.

[151] Casper W. Andersen, Rickard Armiento, Evgeny Blokhin, Gareth J. Conduit, Shyam Dwaraknath, Matthew L. Evans, Ádám Fekete, Abhijith Gopakumar, Saulius Gražulis, Andrius Merkys, Fawzi Mohamed, Corey Oses, Giovanni Pizzi, Gian-Marco Rignanese, Markus Scheidgen, Leopold Talirz, Cormac Toher, Donald Winston, Rossella Aversa, Kamal Choudhary, Pauline Colinet, Stefano Curtarolo, Davide Di Stefano, Claudia Draxl, Suleyman Er, Marco Esters, Marco Fornari, Matteo Giantomassi, Marco Govoni, Geoffroy Hautier, Vinay Hegde, Matthew K. Horton, Patrick Huck, Georg Huhs, Jens Hummelshøj, Ankit Kariryaa, Boris Kozinsky, Snehal Kumbhar, Mohan Liu, Nicola Marzari, Andrew J. Morris, Arash A. Mostofi, Kristin A. Persson, Guido Petretto, Thomas Purcell, Francesco Ricci, Frisco Rose, Matthias Scheffler, Daniel Speckhard, Martin Uhrin, Antanas Vaitkus, Pierre Villars, David Waroquiers, Chris Wolverton, Michael Wu, and Xiaoyu Yang. Optimade, an api for exchanging materials data. *Scientific Data*, 8(1), August 2021.

[152] Robert Arp, Barry Smith, and Andrew D Spear. *Building Ontologies with Basic Formal Ontology*. The MIT Press. MIT Press, London, England, July 2015.

[153] Michel Dumontier, Christopher JO Baker, Joachim Baran, Alison Callahan, Leonid Chepelev, José Cruz-Toledo, Nicholas R Del Rio, Geraint Duck, Laura I Furlong, Nichealla Keath, Dana Klassen, Jamie P McCusker, Núria Queralt-Rosinach, Matthias Samwald, Natalia Villanueva-Rosales, Mark D Wilkinson, and Robert Hoehndorf. The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery. *Journal of Biomedical Semantics*, 5(1), March 2014.

[154] Stefano Borgo, Roberta Ferrario, Aldo Gangemi, Nicola Guarino, Claudio Masolo, Daniele Porello, Emilio M. Sanfilippo, Laure Vieu, Stefano Borgo, Antony Galton, and Oliver Kutz. Dolce: A descriptive ontology for linguistic and cognitive engineering1. *Appl. Ontol.*, 17(1):45–69, January 2022.

[155] et. al. Jesper Friis. emmo-repo/emmo: 1.0.1, 2025.

[156] OWL 2 Web Ontology Language Document Overview (Second Edition) — w3.org. `https://www.w3.org/TR/owl2-overview/`, 2012.

[157] Eric Miller. An introduction to the resource description framework. *Bulletin of the American Society for Information Science and Technology*, 25(1):15–19, October 1998.

[158] Matthew Horridge, Rafael S. Gonçalves, Csongor I. Nyulas, Tania Tudorache, and Mark A. Musen. Webprotégé: A cloud-based ontology editor. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 686–689. ACM, May 2019.

[159] Jean-Baptiste Lamy. Owlready: Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies. *Artificial Intelligence in Medicine*, 80:11–28, July 2017.

[160] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. Pellet: A practical owl-dl reasoner. *Journal of Web Semantics*, 5(2):51–53, June 2007.

[161] Toshihiro Ashino. Materials ontology: An infrastructure for exchanging materials information and knowledge. *Data Science Journal*, 9:54–61, 2010.

[162] J Kaufman and E Begley. Matml: A data interchange markup language, 2003-11-01 00:11:00 2003.

[163] TSP Austin and HH Over. Matdb online - a standards-based system for preserving, managing, and exchanging engineering materials test data. *Data Science Journal*, 11(0):ASMD11–ASMD16, 2012.

[164] Kwok Cheung, John Drennan, and Jane Hunter. Towards an ontology for data-driven discovery of new materials. In *AAAI Spring Symposium: Semantic Scientific Knowledge Integration*, pages 9–14, 2008.

[165] Balashanmuga Priyan Rajamohan, Alexander C. Harding Bradley, Van D. Tran, Jonathan E. Gordon, Hayden W. Caldwell, Redad Mehdi, Gabriel Ponon, Quynh D. Tran, Ozan Dernek, Jarod Kaltenbaugh, Benjamin G. Pierce, Raymond Wieser, Weiqi Yue, Kiefer Lin, Jiana Kambo, Christian Lopez, Arafath Nihar, Daniel J. Savage, Donald W. Brown, Hemant Sharma, Brian Giera, Pawan K. Tripathi, Yinghui Wu, Mengjie Li, Kristopher O. Davis, Laura S. Bruckman, Erika I. Barcelos, and Roger H. French. Materials data science ontology(mds-onto): Unifying domain knowledge in materials and applied data science. *Scientific Data*, 12(1), April 2025.

[166] Jamie P. McCusker, Neha Keshan, Sabbir Rashid, Michael Deagen, Cate Brinson, and Deborah L. McGuinness. *NanoMine: A Knowledge Graph for Nanocomposite Materials Science*, page 144–159. Springer International Publishing, 2020.

[167] Julia Kaltenegger, Kirstine Meyer Frandsen, and Ekaterina Petrova. An ontology-driven framework for digital transformation and performance assessment of building materials. *Building and Environment*, 271:112565, March 2025.

[168] Huanyu Li, Rickard Armiento, and Patrick Lambrix. *An Ontology for the Materials Design Domain*, page 212–227. Springer International Publishing, 2020.

[169] Patrick Lambrix, Rickard Armiento, Huanyu Li, Olaf Hartig, Mina Abd Nikooie Pour, and Ying Li. The materials design ontology. *Semantic Web*, 15(2):481–515, April 2024.

[170] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining: Pearson new international edition*. Pearson custom library. Pearson Education, London, England, July 2013.

[171] Carlos Mera Acosta, Elton Ogoshi, Adalberto Fazzio, Gustavo M. Dalpian, and Alex Zunger. The rashba scale: Emergence of band anti-crossing as a design principle for materials with large rashba coefficient. *Matter*, 3(1):145–165, July 2020.

[172] Ya Zhuo, Aria Mansouri Tehrani, and Jakoah Brgoch. Predicting the band gaps of inorganic solids by machine learning. *The Journal of Physical Chemistry Letters*, 9(7):1668–1673, March 2018.

[173] Olga Kononova, Tanjin He, Haoyan Huo, Amalie Trewartha, Elsa A. Olivetti, and Gerbrand Ceder. Opportunities and challenges of text mining in materials research. *iScience*, 24(3):102155, March 2021.

[174] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, July 2019.

[175] Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. Text-mined dataset of inorganic materials synthesis recipes. *Scientific Data*, 6(1), October 2019.

[176] Callum J. Court and Jacqueline M. Cole. Magnetic and superconducting phase diagrams and transition temperatures predicted using text mining and machine learning. *npj Computational Materials*, 6(1), March 2020.

[177] Callum J. Court, Apoorv Jain, and Jacqueline M. Cole. Inverse design of materials that exhibit the magnetocaloric effect by text-mining of the scientific literature and generative deep learning. *Chemistry of Materials*, 33(18):7217–7231, September 2021.

[178] Amauri J Paula. An automated domain-independent text reading, interpreting and extracting approach for reviewing the scientific literature, 2021.

[179] Lucas Bandeira, Henrique Ferreira, James Moraes de Almeida, Amauri Jardim de Paula, and Gustavo Martini Dalpian. Co2 reduction beyond copper-based catalysts: A natural language processing review from the scientific literature. *ACS Sustainable Chemistry and Engineering*, 12(11):4411–4422, March 2024.

[180] Rishi. et al. Bommasani. On the opportunities and risks of foundation models, 2022.

[181] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[182] Samuel R. Bowman. Eight things to know about large language models, 2023.

[183] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[184] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[185] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[186] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019.

[187] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond, 2023.

[188] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: Pretrained language model for scientific text. In *EMNLP*, 2019.

[189] Tanishq Gupta, Mohd Zaki, N. M. Anoop Krishnan, and Mausam. Matscibert: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1), May 2022.

[190] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori,

Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[191] Microsoft Research AI4Science and Microsoft Azure Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4, 2023.

[192] Maciej P. Polak and Dane Morgan. Extracting accurate materials data from research papers with conversational language models and prompt engineering, 2023.

[193] Zhilong Wang, Yanqiang Han, Xirong Lin, Junfei Cai, Sicheng Wu, and Jinjin Li. An ensemble learning platform for the large-scale exploration of new double perovskites. *ACS Applied Materials; Interfaces*, 14(1):717–725, December 2021.

[194] Carlos Mera Acosta, Elton Ogoshi, Jose Antonio Souza, and Gustavo M. Dalpian. Machine learning study of the magnetic ordering in 2d materials. *ACS Applied Materials; Interfaces*, 14(7):9418–9432, February 2022.

[195] Jarno Laakso, Lauri Himanen, Henrietta Homm, Eiaki V Morooka, Marc OJ Jäger, Milica Todorović, and Patrick Rinke. Updates to the dscribe library: New descriptors and derivatives. *The Journal of Chemical Physics*, 158(23), 2023.

[196] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108(5), January 2012.

[197] Felix Faber, Alexander Lindmaa, O. Anatole von Lilienfeld, and Rickard Armiento. Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry*, 115(16):1094–1101, April 2015.

[198] Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics*, 134(7), February 2011.

[199] Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18), May 2013.

[200] Haoyan Huo and Matthias Rupp. Unified representation of molecules and crystals for machine learning. *Machine Learning: Science and Technology*, 3(4):045017, November 2022.

[201] Mario Valle and Artem R. Oganov. Crystal fingerprint space – a novel paradigm for studying crystal-structure sets. *Acta Crystallographica Section A Foundations of Crystallography*, 66(5):507–517, August 2010.

[202] Seiji Kajita, Nobuko Ohba, Ryosuke Jinnouchi, and Ryoji Asahi. A universal 3d voxel descriptor for solid-state material informatics with deep convolutional neural networks. *Scientific Reports*, 7(1), December 2017.

[203] Prathik R. Kaundinya, Kamal Choudhary, and Surya R. Kalidindi. Prediction of the electron density of states for crystalline compounds with atomistic line graph neural networks (alignn). *JOM*, 74(4):1395–1405, March 2022.

[204] Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, November 2022.

[205] Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Samuel M. Blau, Vlad Cărare, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Fabio Falcioni, Edvin Fako, Andrea C. Ferrari, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Christian Holm, Jad Jaafar, Stephan Hofmann, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Viktor Svahn, Christopher Sutton, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry, 2024.

[206] Runhai Ouyang, Stefano Curtarolo, Emre Ahmetcik, Matthias Scheffler, and Luca M. Ghiringhelli. Sisso: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials*, 2(8), August 2018.

[207] H. Ferreira, E.F. Almeida Junior, W. Espinosa-García, E. Novais, J.N.B. Rodrigues, and G.M. Dalpian. Introduzindo aprendizado de máquina em cursos de física: o caso do rolamento no plano inclinado. *Revista Brasileira de Ensino de Física*, 44, 2022.

[208] Trevor Stephens. GitHub - trevorstephens/gplearn: Genetic Programming in Python, with a scikit-learn inspired API — github.com. `https://github.com/trevorstephens/gplearn`. Acessado em 6 de novembro de 2023.

[209] Silviu-Marian Udrescu and Max Tegmark. Ai feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16), April 2020.

[210] Madhav Muthyala, Farshud Sorourifar, and Joel A. Paulson. Torchsisso: A pytorch-based implementation of the sure independence screening and sparsifying operator for efficient and interpretable model discovery. *Digital Chemical Engineering*, 13:100198, December 2024.

[211] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

[212] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc.

[213] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, January 2015.

[214] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5), October 2001.

[215] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144. ACM, August 2016.

[216] Joohwi Lee and Kaito Miyamoto. Accurate predictive model of band gap with selected important features based on explainable machine learning, 2025.

[217] Taichi Masuda and Katsuaki Tanabe. Explainable artificial intelligence for machine learning prediction of bandgap energies. *Journal of Applied Physics*, 136(17), November 2024.

[218] Elton Ogoshi, Mário Popolin-Neto, Carlos Mera Acosta, Gabriel M. Nascimento, João N. B. Rodrigues, Osvaldo N. Oliveira, Fernando V. Paulovich, and Gustavo M. Dalpian. Learning from machine learning: the case of band-gap directness in semiconductors. *Discover Materials*, 4(1), February 2024.

[219] M. Hutchinson and M. Widom. Vasp on a gpu: Application to exact-exchange calculations of the stability of elemental boron. *Computer Physics Communications*, 183(7):1422–1426, July 2012.

[220] Tao Wang, Xiaoxing Tan, Yadong Wei, and Hao Jin. Accurate bandgap predictions of solids assisted by machine learning. *Materials Today Communications*, 29:102932, December 2021.

[221] Pedro A. Santos-Florez, Howard Yanxon, Byungkyun Kang, Yansun Yao, and Qiang Zhu. Size-dependent nucleation in crystal phase transition from machine learning metadynamics. *Physical Review Letters*, 129(18), October 2022.

[222] Dylan M. Anstine and Olexandr Isayev. Machine learning interatomic potentials and long-range physics. *The Journal of Physical Chemistry A*, 127(11):2417–2431, February 2023.

[223] Artem R. Oganov. Crystal structure prediction: reflections on present status and challenges. *Faraday Discussions*, 211:643–660, 2018.

[224] Daniel W. Davies, Keith T. Butler, Adam J. Jackson, Andrew Morris, Jarvist M. Frost, Jonathan M. Skelton, and Aron Walsh. Computational screening of all stoichiometric inorganic materials. *Chem*, 1(4):617–627, October 2016.

[225] Colin W. Glass, Artem R. Oganov, and Nikolaus Hansen. Uspex—evolutionary crystal structure prediction. *Computer Physics Communications*, 175(11–12):713–720, December 2006.

[226] Paul C. Jennings, Steen Lysgaard, Jens Strabo Hummelshøj, Tejs Vegge, and Thomas Bligaard. Genetic algorithms for computational materials discovery accelerated by machine learning. *npj Computational Materials*, 5(1), April 2019.

[227] Callum J. Court, Batuhan Yildirim, Apoorv Jain, and Jacqueline M. Cole. 3-d inorganic crystal structure generation and property prediction via representation learning. *Journal of Chemical Information and Modeling*, 60(10):4518–4535, August 2020.

[228] Zhenpeng Yao, Benjamín Sánchez-Lengeling, N. Scott Bobbitt, Benjamin J. Bucior, Sai Govind Hari Kumar, Sean P. Collins, Thomas Burns, Tom K. Woo, Omar K. Farha, Randall Q. Snurr, and Alán Aspuru-Guzik. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nature Machine Intelligence*, 3(1):76–86, January 2021.

[229] Cai-Yuan Ye, Hong-Ming Weng, and Quan-Sheng Wu. Con-cdvae: A method for the conditional generation of crystal structures. *Computational Materials Today*, 1:100003, May 2024.

[230] Microsoft Research AI4Science and Microsoft Azure Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4, 2023.

[231] Hyunsoo Park, Zhenzhu Li, and Aron Walsh. Has generative artificial intelligence solved inverse materials design? *Matter*, 7(7):2355–2367, July 2024.

[232] Edirisuriya M. Dilanga Siriwardane, Yong Zhao, Indika Perera, and Jianjun Hu. Generative design of stable semiconductor materials using deep learning and density functional theory. *npj Computational Materials*, 8(1), August 2022.

[233] Sherry Yang, KwangHwan Cho, Amil Merchant, Pieter Abbeel, Dale Schuurmans, Igor Mordatch, and Ekin Dogus Cubuk. Scalable diffusion for materials generation, 2023.

[234] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, Roberto Sordillo, Lixin Sun, Jake Smith, Bichlien Nguyen, Hannes Schulz, Sarah Lewis, Chin-Wei Huang, Ziheng Lu, Yichi Zhou, Han Yang, Hongxia Hao, Jielan Li, Chunlei Yang, Wenjie Li, Ryota Tomioka, and Tian Xie. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, January 2025.

[235] Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, November 2023.

[236] Jeffrey Travis and Jim Kring. *LabVIEW for everyone: Graphical programming made easy and fun*. Prentice Hall PTR, Harlow, England, 2007.

[237] Ricardo Macarron, Martyn N. Banks, Dejan Bojanic, David J. Burns, Dragan A. Cirovic, Tina Garyantes, Darren V. S. Green, Robert P. Hertzberg, William P. Janzen, Jeff W. Paslay, Ulrich Schopfer, and G. Sitta Sittampalam. Impact of high-throughput screening in biomedical research. *Nature Reviews Drug Discovery*, 10(3):188–195, March 2011.

[238] Aldair E. Gongora, Bowen Xu, Wyatt Perry, Chika Okoye, Patrick Riley, Kristofer G. Reyes, Elise F. Morgan, and Keith A. Brown. A bayesian experimental autonomous researcher for mechanical design. *Science Advances*, 6(15), April 2020.

[239] Aldair E. Gongora, Kelsey L. Snapp, Emily Whiting, Patrick Riley, Kristofer G. Reyes, Elise F. Morgan, and Keith A. Brown. Using simulation to accelerate autonomous experimentation: A case study using mechanics. *iScience*, 24(4):102262, April 2021.

[240] Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990):86–91, November 2023.

[241] Wei Li, Zheming Wang, Felix Deschler, Song Gao, Richard H. Friend, and Anthony K. Cheetham. Chemically diverse and multifunctional hybrid organic–inorganic perovskites. *Nature Reviews Materials*, 2(3), February 2017.

[242] David A. Egger, Andrew M. Rappe, and Leeor Kronik. Hybrid organic–inorganic perovskites on the move. *Accounts of Chemical Research*, 49(3):573–581, February 2016.

[243] Shiro Kawachi, Mika Atsumi, Noriko Saito, Naoki Ohashi, Youichi Murakami, and Jun-ichi Yamaura. Structural and thermal properties in formamidinium and cs-mixed

lead halides. *The Journal of Physical Chemistry Letters*, 10(22):6967–6972, October 2019.

[244] Quinten A. Akkerman and Liberato Manna. What defines a halide perovskite? *ACS Energy Letters*, 5(2):604–610, January 2020.

[245] M. I. Balanov, O. A. Brylev, R. Kevorkyants, A. V. Emeline, N. I. Selivanov, Y. V. Chizhov, A. V. Syuy, D. S. Shtarev, and V. V. Korochencev. Crystal structure and electronic properties of low-dimensional hexamethylenediaminium lead halide perovskites. *Dalton Transactions*, 52(19):6388–6397, 2023.

[246] Giulia Grancini and Mohammad Khaja Nazeeruddin. Dimensional tailoring of hybrid perovskites for photovoltaics. *Nature Reviews Materials*, 4(1):4–22, November 2018.

[247] Constantinos C. Stoumpos, Duyen H. Cao, Daniel J. Clark, Joshua Young, James M. Rondinelli, Joon I. Jang, Joseph T. Hupp, and Mercouri G. Kanatzidis. Ruddlesden–popper hybrid lead iodide perovskite 2d homologous semiconductors. *Chemistry of Materials*, 28(8):2852–2867, April 2016.

[248] Antonio Gaetano Ricciardulli, Sheng Yang, Jurgen H. Smet, and Michael Saliba. Emerging perovskite monolayers. *Nature Materials*, 20(10):1325–1336, June 2021.

[249] G. A. Mousdis, G. C. Papavassiliou, C. P. Raptopoulou, and A. Terzis. Preparation and characterization of [h3n(ch2)6nh3]pbi4 and similar compounds with a layered perovskite structure. *Journal of Materials Chemistry*, 10(2):515–518, 2000.

[250] D. B. Mitzi, C. A. Feild, W. T. A. Harrison, and A. M. Guloy. Conducting tin halides with a layered organic-based perovskite structure. *Nature*, 369(6480):467–469, June 1994.

[251] Rui Su, Antonio Fieramosca, Qing Zhang, Hai Son Nguyen, Emmanuelle Deleporte, Zhanghai Chen, Daniele Sanvitto, Timothy C. H. Liew, and Qihua Xiong. Perovskite semiconductors for room-temperature exciton-polaritonics. *Nature Materials*, 20(10):1315–1324, July 2021.

[252] Chan Myae Myae Soe, Constantinos C Stoumpos, Mikaël Kepenekian, Boubacar Traoré, Hsinhan Tsai, Wanyi Nie, Binghao Wang, Claudine Katan, Ram Seshadri, Aditya D Mohite, Jacky Even, Tobin J Marks, and Mercouri G Kanatzidis. New type of 2D perovskites with alternating cations in the interlayer space, $(C(NH_2)_3)(CH_3NH_3)_nPb_nI_{3n+1}$: Structure, properties, and photovoltaic performance. *J. Am. Chem. Soc.*, 139(45):16297–16309, November 2017.

[253] Xiaotong Li, Justin M. Hoffman, and Mercouri G. Kanatzidis. The 2d halide perovskite rulebook: How the spacer influences everything from the structure to optoelectronic device efficiency. *Chemical Reviews*, 121(4):2230–2291, January 2021.

[254] Zewen Xiao, Weiwei Meng, Jianbo Wang, David B. Mitzi, and Yanfa Yan. Searching for promising new perovskite-based photovoltaic absorbers: the importance of electronic dimensionality. *Materials Horizons*, 4(2):206–216, 2017.

[255] Bin Cheng, Ting-You Li, Partha Maity, Pai-Chun Wei, Dennis Nordlund, Kang-Ting Ho, Der-Hsien Lien, Chun-Ho Lin, Ru-Ze Liang, Xiaohe Miao, Idris A. Ajia, Jun Yin, Dimosthenis Sokaras, Ali Javey, Iman S. Roqan, Omar F. Mohammed, and Jr-Hau He. Extremely reduced dielectric confinement in two-dimensional hybrid perovskites with large polar organics. *Communications Physics*, 1(1), November 2018.

[256] Rayan Chakraborty and Angshuman Nag. Correlation of dielectric confinement and excitonic binding energy in 2d layered hybrid perovskites using temperature dependent photoluminescence. *The Journal of Physical Chemistry C*, 124(29):16177–16185, June 2020.

[257] Guangda Niu, Xudong Guo, and Liduo Wang. Review of recent progress in chemical stability of perovskite solar cells. *Journal of Materials Chemistry A*, 3(17):8970–8980, 2015.

[258] Matthew D. Smith, Laurent Pedesseau, Mikaël Kepenekian, Ian C. Smith, Claudine Katan, Jacky Even, and Hemamala I. Karunadasa. Decreasing the electronic confinement in layered perovskites through intercalation. *Chemical Science*, 8(3):1960–1968, 2017.

[259] V. M. Goldschmidt. Die gesetze der krystallochemie. *Die Naturwissenschaften*, 14(21):477–485, May 1926.

[260] Marina R. Filip and Feliciano Giustino. The geometric blueprint of perovskites. *Proceedings of the National Academy of Sciences*, 115(21):5397–5402, May 2018.

[261] Christopher J. Bartel, Christopher Sutton, Bryan R. Goldsmith, Runhai Ouyang, Charles B. Musgrave, Luca M. Ghiringhelli, and Matthias Scheffler. New tolerance factor to predict the stability of perovskite oxides and halides. *Science Advances*, 5(2), February 2019.

[262] C. Meric Guvenc, Stefano Toso, Yurii P. Ivanov, Gabriele Saleh, Sinan Balci, Giorgio Divitini, and Liberato Manna. Breaking the boundaries of the goldschmidt tolerance factor with ethylammonium lead iodide perovskite nanocrystals. *ACS Nano*, 19(1):1557–1565, December 2024.

[263] Lung-Chien Chen, Kuan-Lin Lee, and Shao-En Lin. Observation of hybrid mapbbr3 perovskite bulk crystals grown by repeated crystallizations. *Crystals*, 8(7):260, June 2018.

[264] Lin-Jie Yang, Wenye Xuan, Sara Henda, Shaoyang Wang, Sai Kiran Rajendran, David B. Cordes, David N. Miller, Alexandra M. Z. Slawin, Lethy Krishnan Jagadamma, Hamid Ohadi, Hsin-Yi Tiffany Chen, Matthew S. Dyer, and Julia L. Payne. Advancing intercalation strategies in layered hybrid perovskites by bringing together synthesis and simulations. *Nature Communications*, 16(1), July 2025.

[265] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[266] J. Tauc. Optical properties and electronic structure of amorphous ge and si. *Materials Research Bulletin*, 3(1):37–46, January 1968.

[267] Franz Urbach. The long-wavelength edge of photographic sensitivity and of the electronic absorption of solids. *Physical Review*, 92(5):1324–1324, December 1953.

[268] Alex M Ganose, Adam J Jackson, and David O Scanlon. sumo: Command-line tools for plotting and analysis of periodic ab initio calculations. *Journal of Open Source Software*, 3(28):717, August 2018.

[269] Ruo Xi Yang, Matthew K. Horton, Jason Munro, and Kristin A. Persson. High-throughput optical absorption spectra for inorganic semiconductors, 2022.

[270] Bikash Patra, Subrata Jana, Lucian A. Constantin, and Prasanjit Samal. Efficient band gap prediction of semiconductors and insulators from a semilocal exchange-correlation functional. *Physical Review B*, 100(4), July 2019.

[271] Eliane A. Morais, Naidel A. M. S. Caturello, Maykon A. Lemes, Henrique Ferreira, Fabio F. Ferreira, Jose J. S. Acuña, Sergio Brochsztain, Gustavo M. Dalpian, and Jose A. Souza. Rashba spin splitting limiting the application of 2d halide perovskites for uv-emitting devices. *ACS Applied Materials & Interfaces*, 16(3):4261–4270, January 2024.

[272] Menglin Li, Peng Huang, and Haizheng Zhong. Current understanding of band-edge properties of halide perovskites: Urbach tail, rashba splitting, and exciton binding energy. *The Journal of Physical Chemistry Letters*, 14(6):1592–1603, February 2023.

[273] Mariusz Wojcik and M. Tachiya. Accuracies of the empirical theories of the escape probability based on eigen model and braun model compared with the exact extension of onsager theory. *The Journal of Chemical Physics*, 130(10), March 2009.

[274] Kanwar S. Nalwa, John A. Carr, Rakesh C. Mahadevapuram, Hari K. Kodali, Sayantan Bose, Yuqing Chen, Jacob W. Petrich, Baskar Ganapathysubramanian, and Sumit Chaudhary. Enhanced charge separation in organic photovoltaic films doped with ferroelectric dipoles. *Energy &; Environmental Science*, 5(5):7042, 2012.

[275] T. Balić-Žunić, A. Garavelli, D. Mitolo, P. Acquafredda, and E. Leonardsen. Jakobssonite, CaAlF5, a new mineral from fumaroles at the eldfell and hekla volcanoes, iceland. *Mineralogical Magazine*, 76(3):751–760, June 2012.

[276] S.P. Jakobsson, E.S. Leonardsen, Tonci Balic Zunic, and Jónsson S.S. *Encrustations from three recent volcanic eruptions in Iceland: the 1963-1967 Surtsey, the 1973 Eldfell and the 1991 Hekla eruptions*, volume 52. Fjölrit Náttúrufraedistofnunar, 2008.

[277] Jan Lützow Holm, Torsten Johnsson, Terje Skansen, J. Brunvoll, and Merv Hinton. Phase equilibria in the system CaF2-AlF3. *Acta Chemica Scandinavica*, 19:1512–1514, 1965.

[278] Alexander Rehmer, Kerstin Scheurell, Gudrun Scholz, and Erhard Kemnitz. Sol-gel-synthesis of nanoscopic complex metal fluorides. *Nanomaterials*, 7(11):362, November 2017.

[279] D. Dumora, R. Von der Mühll, and J. Ravez. Structure cristalline de CaCrF5. *Materials Research Bulletin*, 6(7):561–569, July 1971.

[280] Kang Kun Wu and I.D. Brown. Refinement of the crystal structure of CaCrF5. *Materials Research Bulletin*, 8(5):593–598, May 1973.

[281] Régnault von der Mühll, Françoise Daut, and Jean Ravez. Structure cristalline de SrFeF5. *Journal of Solid State Chemistry*, 8(3):206–212, November 1973.

[282] Rainer Domesle and Rudolf Hoppe. The crystal structure of ca2alf7. *Zeitschrift für Kristallographie*, 153(3-4):317–328, February 1980.

[283] Paul Hagenmuller, editor. *Inorganic Solid Fluorides*. Elsevier, 1985.

[284] A. Hemon and G. Courbion. Refinement of a room-temperature structure of α-$CaAlF_5$. *Acta Crystallographica Section C*, 47(6):1302–1303, June 1991.

[285] Tobias Schlöder, Florian Kraus, and Sebastian Riedel. *Fluorides: Solid-State Chemistry*, pages 1–30. American Cancer Society, 2014.

[286] M. Body, G. Silly, C. Legein, J.-Y. Buzaré, F. Calvayrac, and P. Blaha. Structural investigations of β-$CaAlF_5$ by coupling powder XRD, NMR, EPR and spectroscopic parameter calculations. *Journal of Solid State Chemistry*, 178(12):3655–3661, December 2005.

[287] L.A. Curtiss and A. Heinricher. Ab initio molecular orbital studies of the structures and potential energy surfaces of the bealf5 and mgalf5 complexes. *Chemical Physics Letters*, 86(5):467–471, 1982.

[288] Laurent Joubert, Gérard Picard, Bernard Silvi, and Franck Fuster. Electron localization function view of bonding in selected aluminum fluoride molecules. *Journal of Molecular Structure: THEOCHEM*, 463(1-2):75–80, 1999.

[289] Hirotsugu Inoue, Tokuro Nanba, H. Hagihara, T. Kanazawa, and Itaru Yasui. Computer simulation of raman spectra of fluoride glasses. In *Halide Glasses V*, volume 32 of *Materials Science Forum*, pages 403–408. Trans Tech Publications Ltd, 1 1991.

[290] G. Kresse and J. Hafner. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B*, 47:558–561, Jan 1993.

[291] Yury V. Kissin, Thomas E. Nowlin, and Robert I. Mink. Supported titanium/magnesium ziegler catalysts for the production of polyethylene, April 2018.

[292] D.K. Ingole, C.P. Joshi, S.V. Moharil, P.L. Muthal, and S.M. Dhopte. Luminescence of ce3+ in some chloroaluminates. *Journal of Luminescence*, 130(7):1194–1197, July 2010.

[293] Ying Han, Jie Zhou, Tieqiang Li, Zheng Yi, Ning Lin, and Yitai Qian. Molten-salt chemical exfoliation process for preparing two-dimensional mesoporous si nanosheets as high-rate li-storage anode. *Nano Research*, 11(12):6294–6303, August 2018.

[294] Pavel P Fedorov, Ruslan M Zakalyukin, Lidiya N Ignat'eva, and Vyacheslav M Bouznik. Fluoroindate glasses. *Russian Chemical Reviews*, 69(8):705–716, August 2000.

[295] Matthias Weil, Erich Zobetz, Franz Werner, and Frank Kubel. New alkaline earth aluminium fluorides with the formula $(\mathrm{M}, \mathrm{M}`)\mathrm{AlF}_5$ $(\mathrm{M}, \mathrm{M}` = \mathrm{Ca}, \mathrm{Sr}, \mathrm{Ba})$. *Solid State Sciences*, 3(4):441–453, May 2001.

[296] J. Bardeen and W. H. Brattain. The transistor, a semi-conductor triode. *Physical Review*, 74(2):230–231, July 1948.

[297] Shuji Nakamura, Masayuki Senoh, Naruhito Iwasa, and Shin-ichi Nagahama Shin-ichi Nagahama. High-brightness ingan blue, green and yellow light-emitting diodes with quantum well structures. *Japanese Journal of Applied Physics*, 34(7A):L797, July 1995.

[298] J. Y. Tsao, S. Chowdhury, M. A. Hollis, D. Jena, N. M. Johnson, K. A. Jones, R. J. Kaplar, S. Rajan, C. G. Van de Walle, E. Bellotti, C. L. Chua, R. Collazo, M. E. Coltrin, J. A. Cooper, K. R. Evans, S. Graham, T. A. Grotjohn, E. R. Heller, M. Higashiwaki, M. S. Islam, P. W. Juodawlkis, M. A. Khan, A. D. Koehler, J. H. Leach, U. K. Mishra, R. J. Nemanich, R. C. N. Pilawa-Podgurski, J. B. Shealy, Z. Sitar,

M. J. Tadjer, A. F. Witulski, M. Wraback, and J. A. Simmons. Ultrawide-bandgap semiconductors: Research opportunities and challenges. *Advanced Electronic Materials*, 4(1), December 2017.

[299] Masataka Higashiwaki, Robert Kaplar, Julien Pernot, and Hongping Zhao. Ultrawide bandgap semiconductors. *Applied Physics Letters*, 118(20), May 2021.

[300] Man Hoi Wong, Oliver Bierwagen, Robert J. Kaplar, and Hitoshi Umezawa. Ultrawide-bandgap semiconductors: An overview. *Journal of Materials Research*, 36(23):4601–4615, December 2021.

[301] T C McGill and D A Collins. Prospects for the future of narrow bandgap materials. *Semiconductor Science and Technology*, 8(1S):S1–S5, January 1993.

[302] Xiao-Hui Li, Yi-Xuan Guo, Yujie Ren, Jia-Jun Peng, Ji-Shu Liu, Cong Wang, and Han Zhang. Narrow-bandgap materials for optoelectronics applications. *Frontiers of Physics*, 17(1), July 2021.

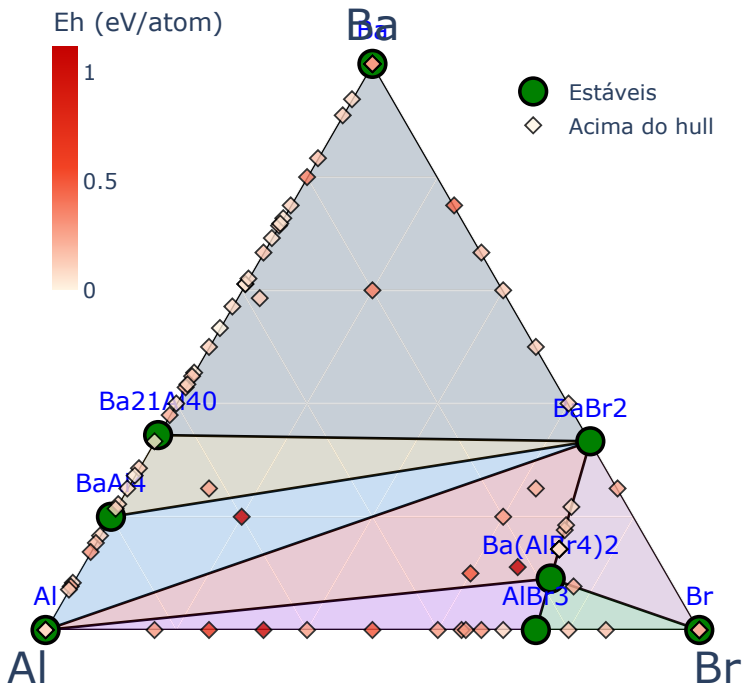
[303] Łukasz Mentel. mendeleev – a python resource for properties of chemical elements, ions and isotopes.

[304] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Apêndice A

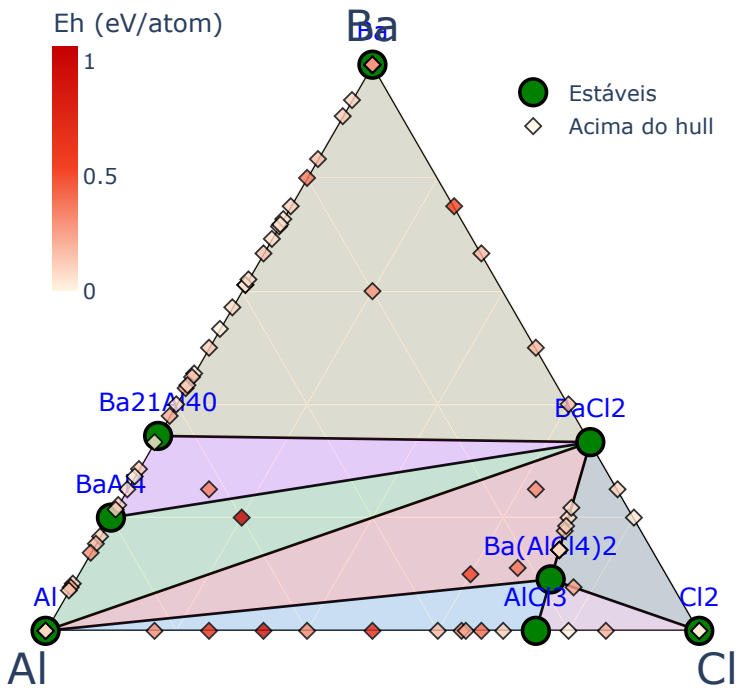
Neste apêndice são apresentados os diagramas de fase para cada sistema simulado por HT-DFT referentes ao estudo de caso II, discutido no Capítulo 5. Esses diagramas de fase mostram os materiais estáveis com relação a energia de suas fases competidoras - distância E_h em energia do respectivo material para a superfície formada pelas estruturas de mínima energia, também denominado *convex hull*. Uma discussão dos conceitos de fases competidoras e estabilidade termodinâmica é feita no Capítulo 2.

Figura A1: Diagrama de fase do sistema Ba-Al-Br.



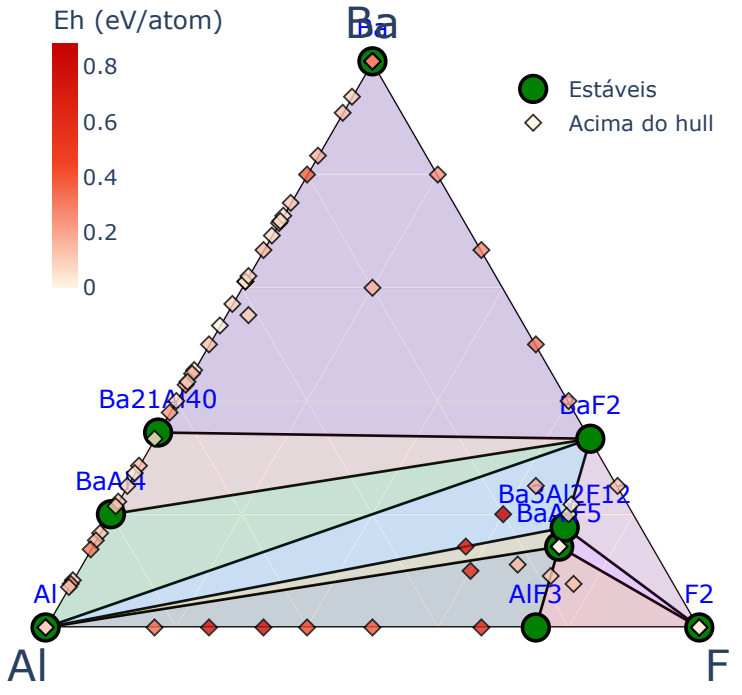
Fonte: elaborado pelo autor.

Figura A2: Diagrama de fase do sistema Ba-Al-Cl.



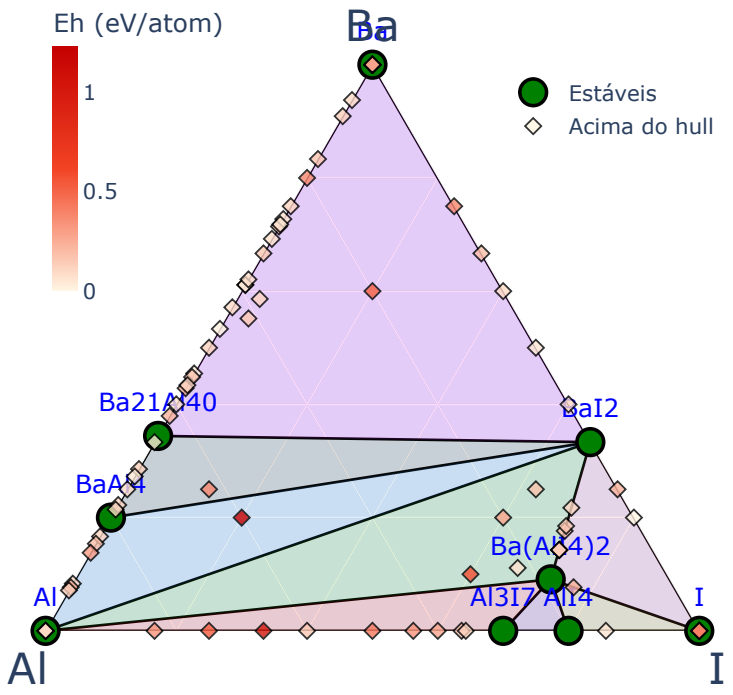
Fonte: elaborado pelo autor.

Figura A3: Diagrama de fase do sistema Ba-Al-F.



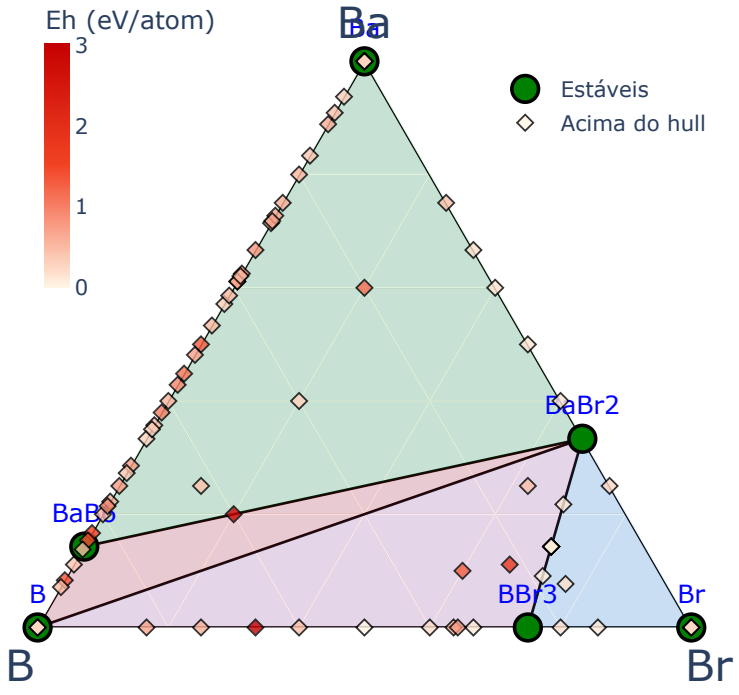
Fonte: elaborado pelo autor.

Figura A4: Diagrama de fase do sistema Ba-Al-I.



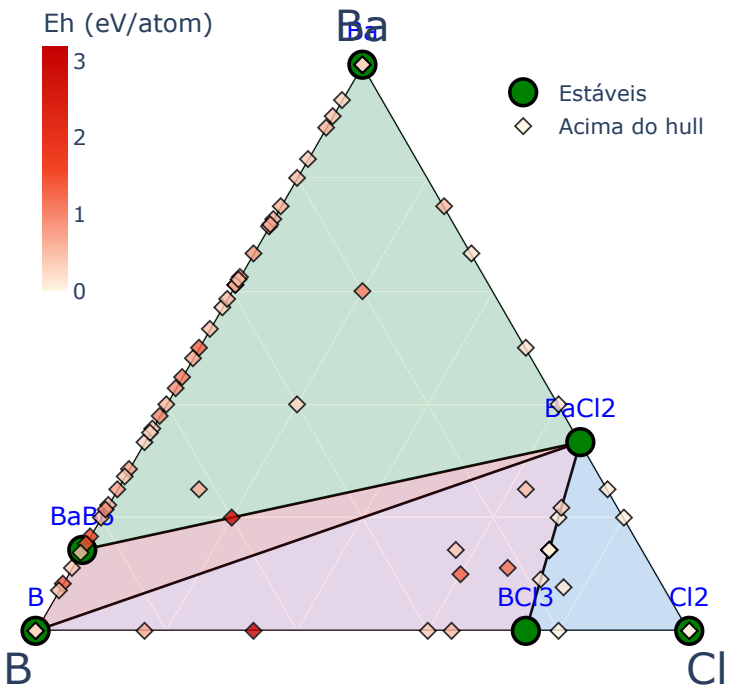
Fonte: elaborado pelo autor.

Figura A5: Diagrama de fase do sistema Ba-B-Br.



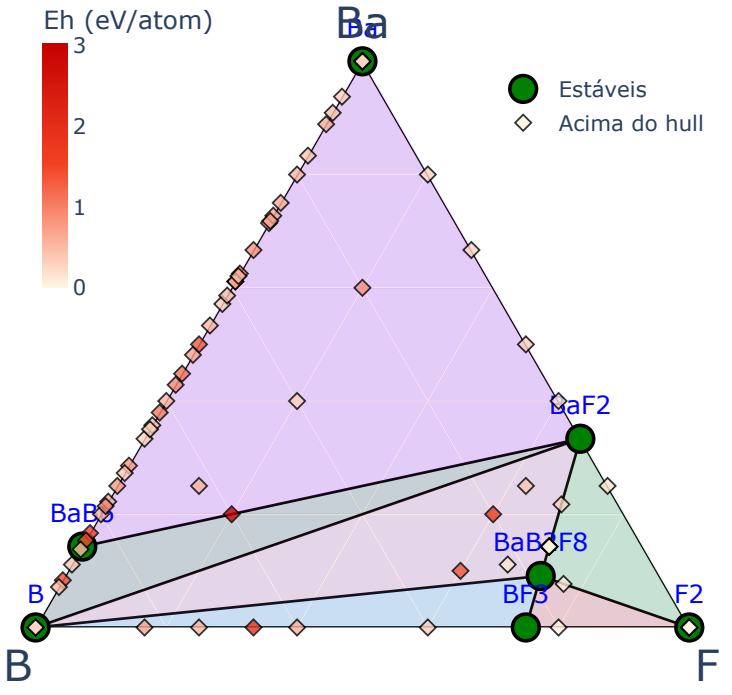
Fonte: elaborado pelo autor.

Figura A6: Diagrama de fase do sistema Ba-B-Cl.



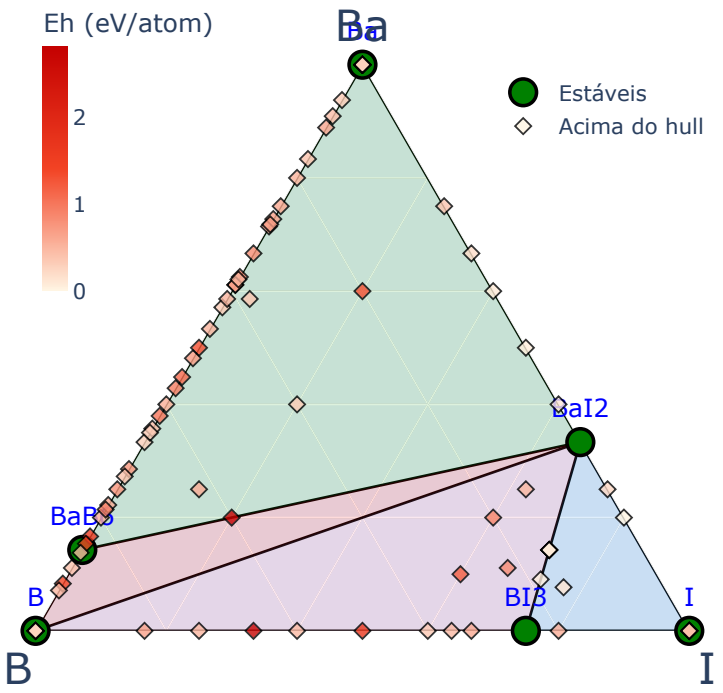
Fonte: elaborado pelo autor.

Figura A7: Diagrama de fase do sistema Ba-B-F.



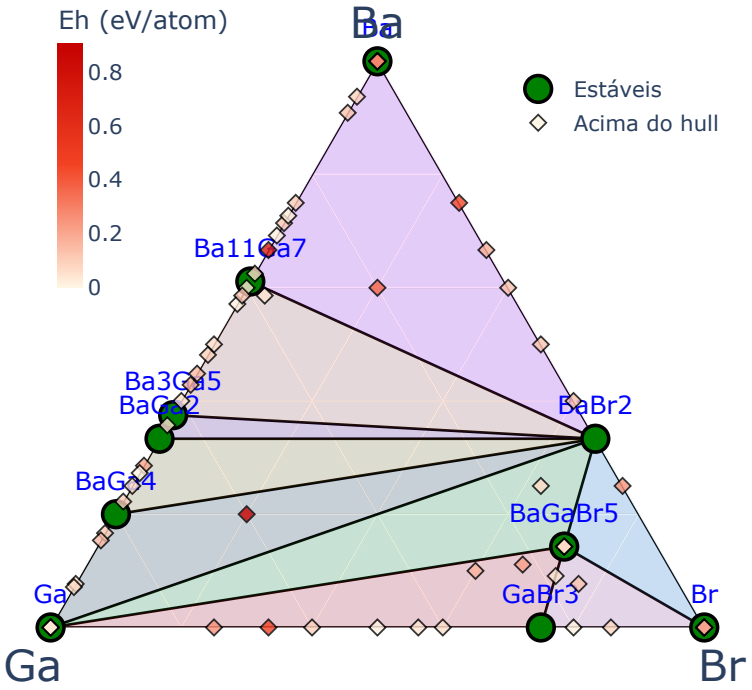
Fonte: elaborado pelo autor.

Figura A8: Diagrama de fase do sistema Ba-B-I.



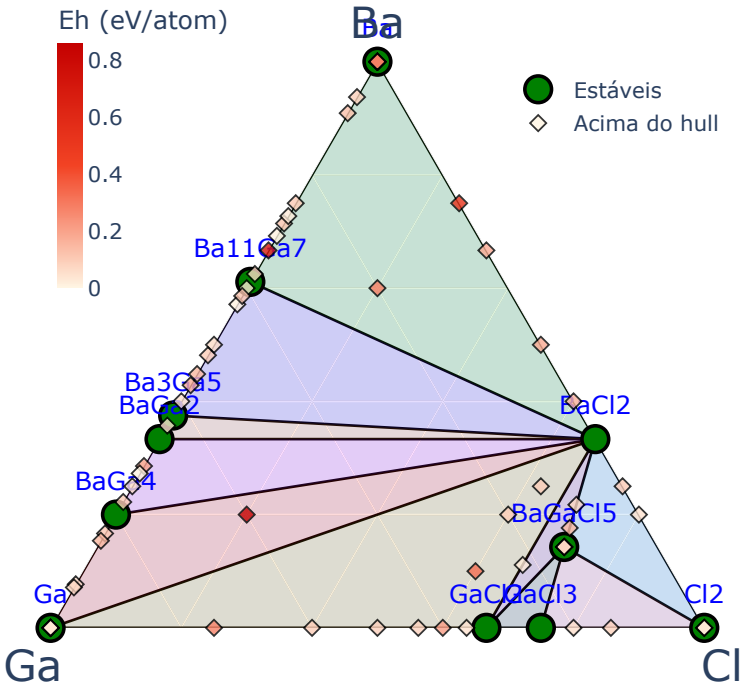
Fonte: elaborado pelo autor.

Figura A9: Diagrama de fase do sistema Ba-Ga-Br.



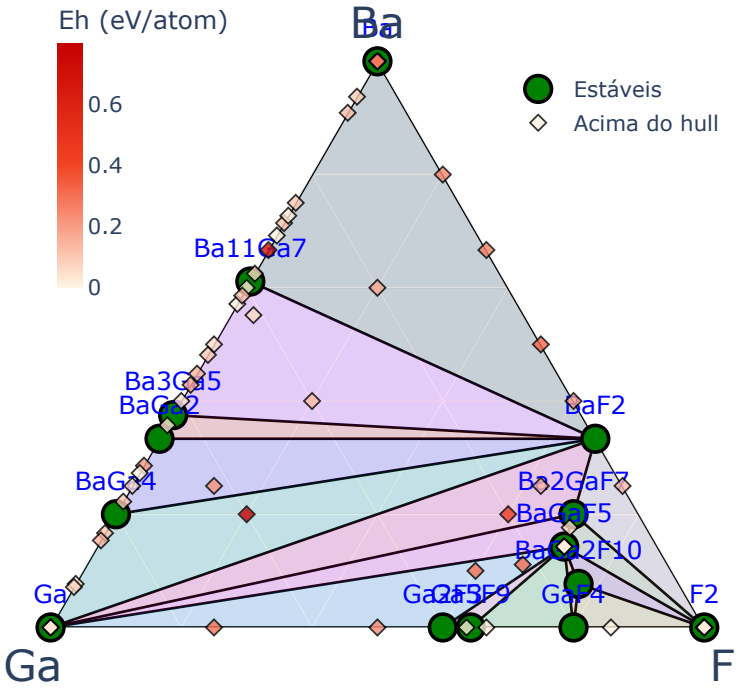
Fonte: elaborado pelo autor.

Figura A10: Diagrama de fase do sistema Ba-Ga-Cl.



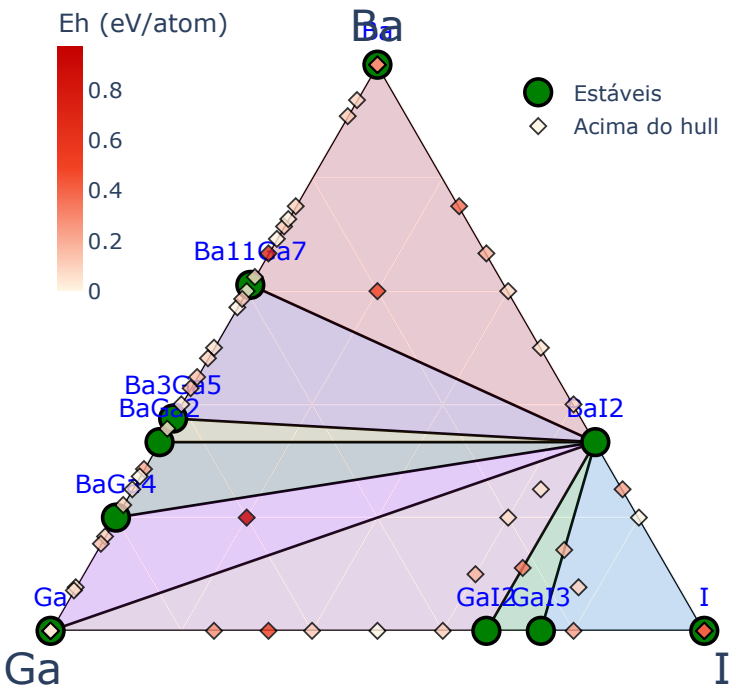
Fonte: elaborado pelo autor.

Figura A11: Diagrama de fase do sistema Ba-Ga-F.



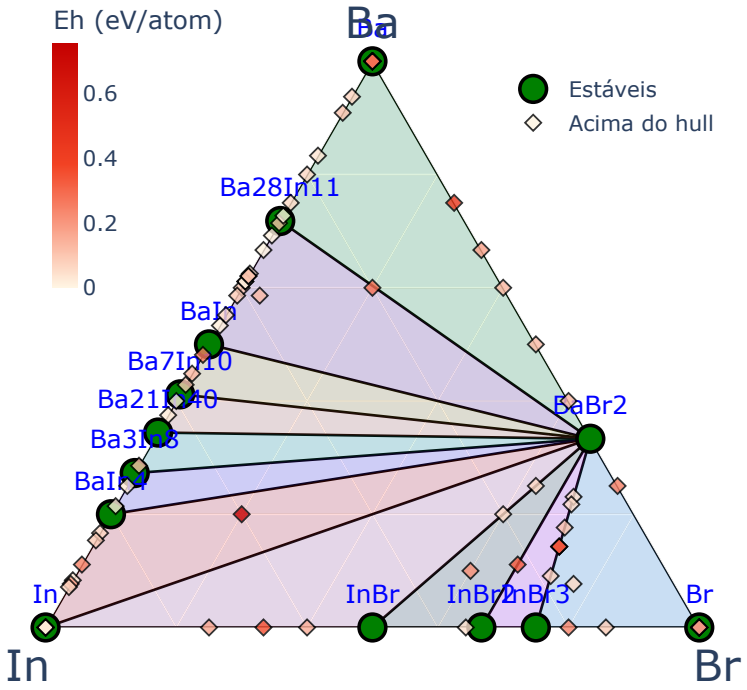
Fonte: elaborado pelo autor.

Figura A12: Diagrama de fase do sistema Ba-Ga-I.



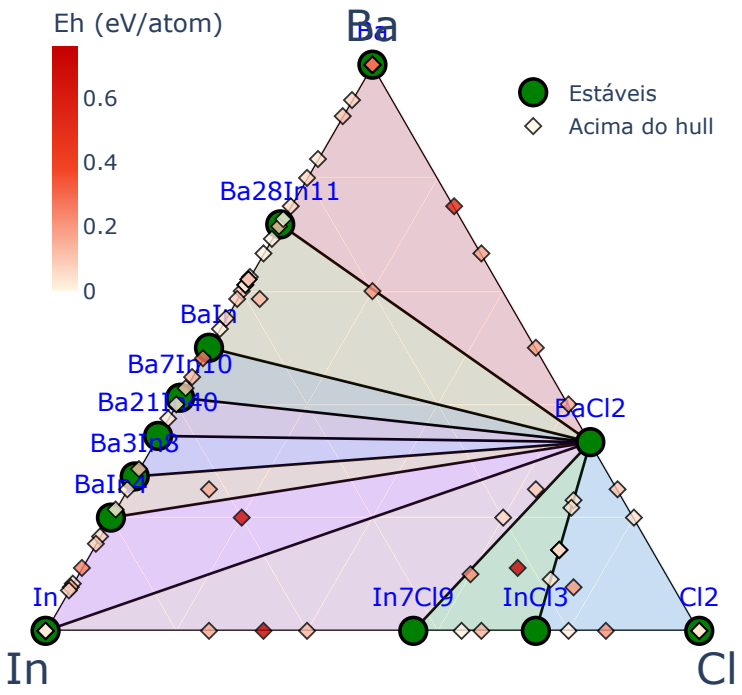
Fonte: elaborado pelo autor.

Figura A13: Diagrama de fase do sistema Ba-In-Br.



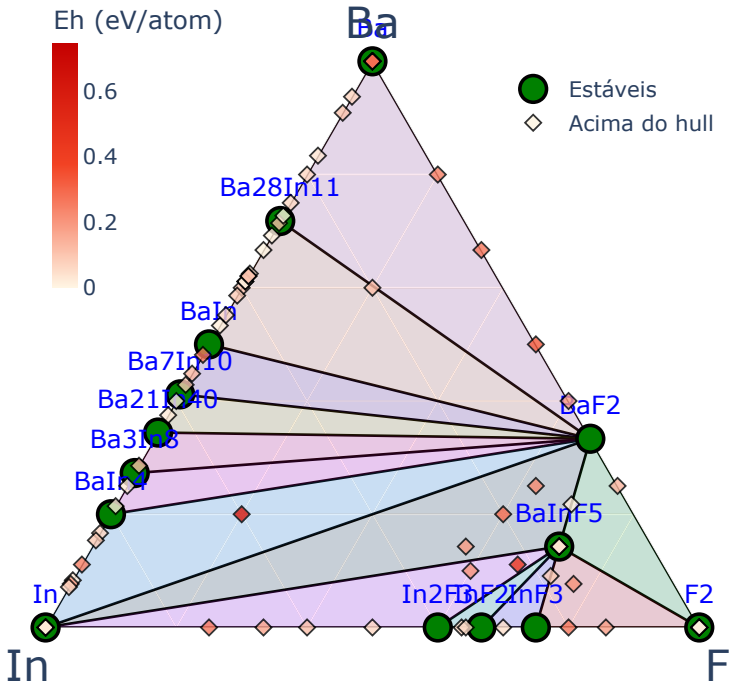
Fonte: elaborado pelo autor.

Figura A14: Diagrama de fase do sistema Ba-In-Cl.



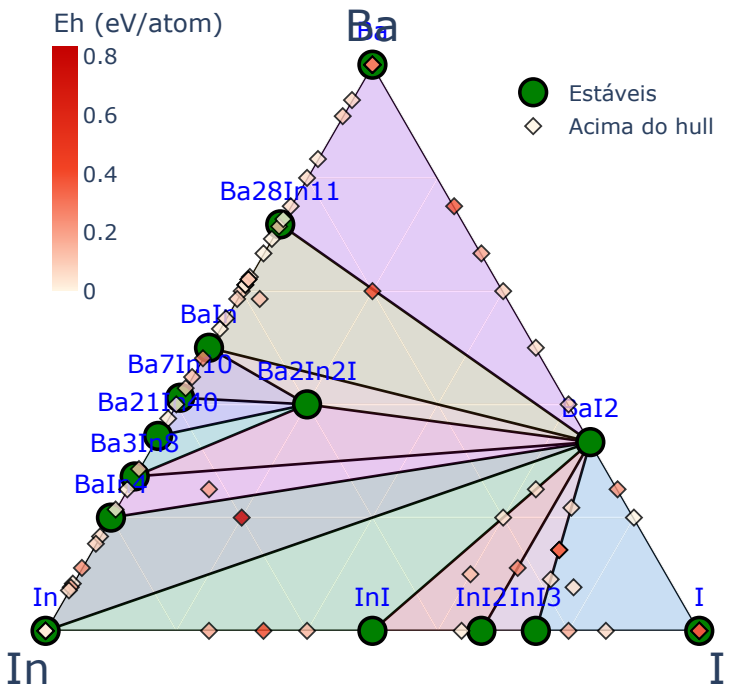
Fonte: elaborado pelo autor.

Figura A15: Diagrama de fase do sistema Ba-In-F.



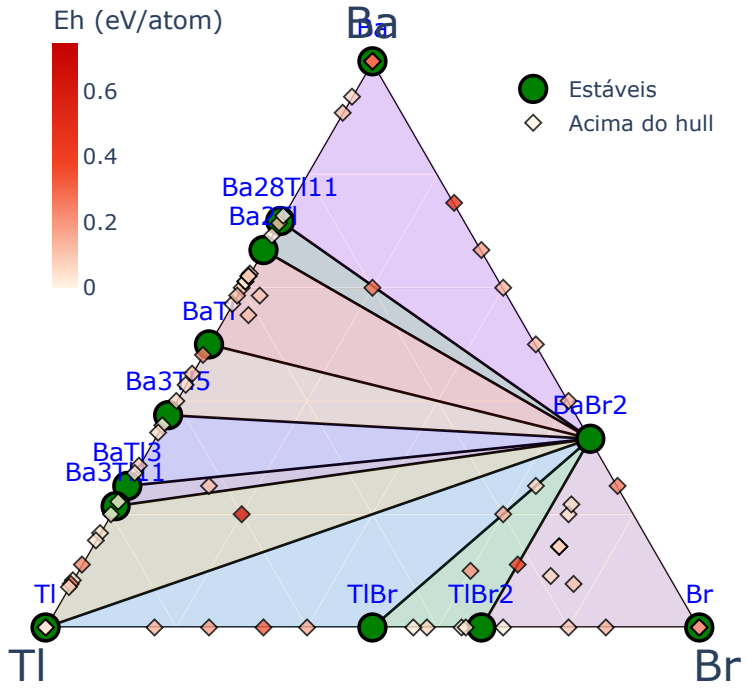
Fonte: elaborado pelo autor.

Figura A16: Diagrama de fase do sistema Ba-In-I.



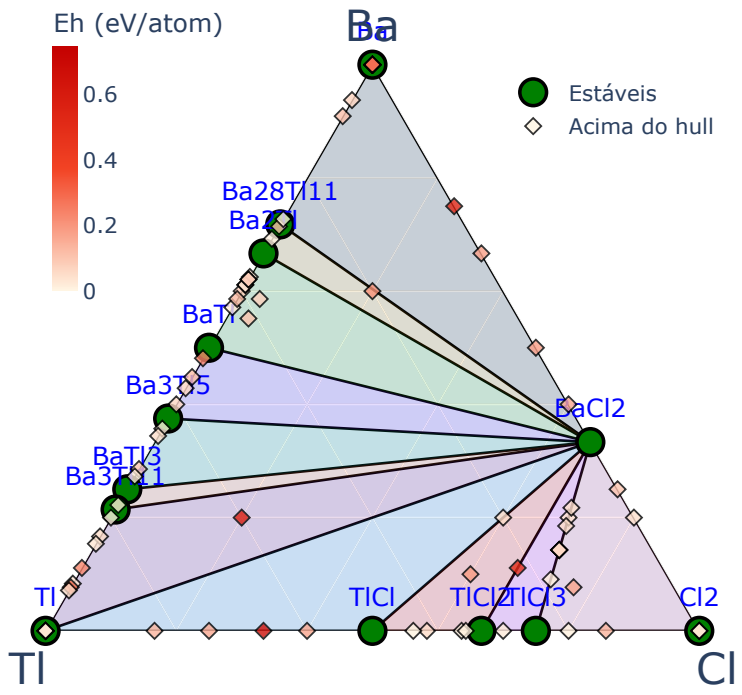
Fonte: elaborado pelo autor.

Figura A17: Diagrama de fase do sistema Ba-Tl-Br.



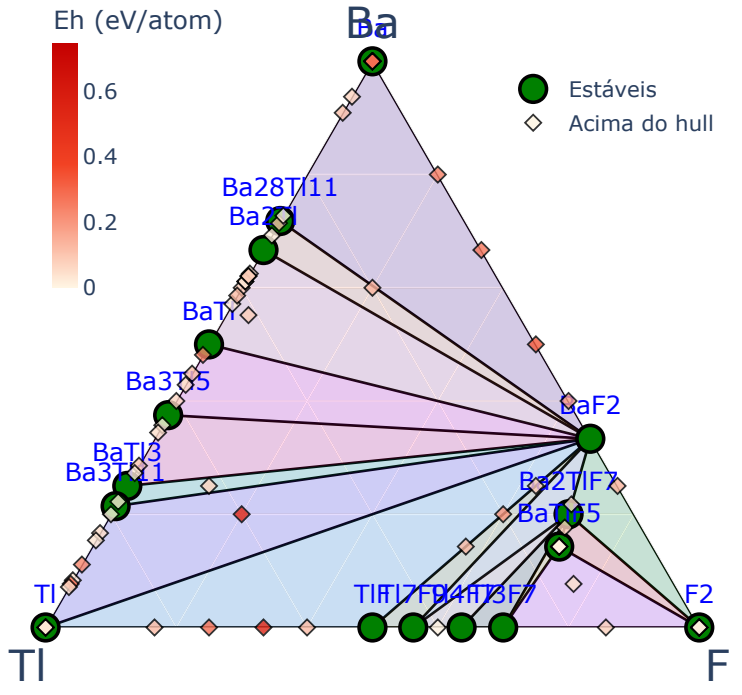
Fonte: elaborado pelo autor.

Figura A18: Diagrama de fase do sistema Ba-Tl-Cl.



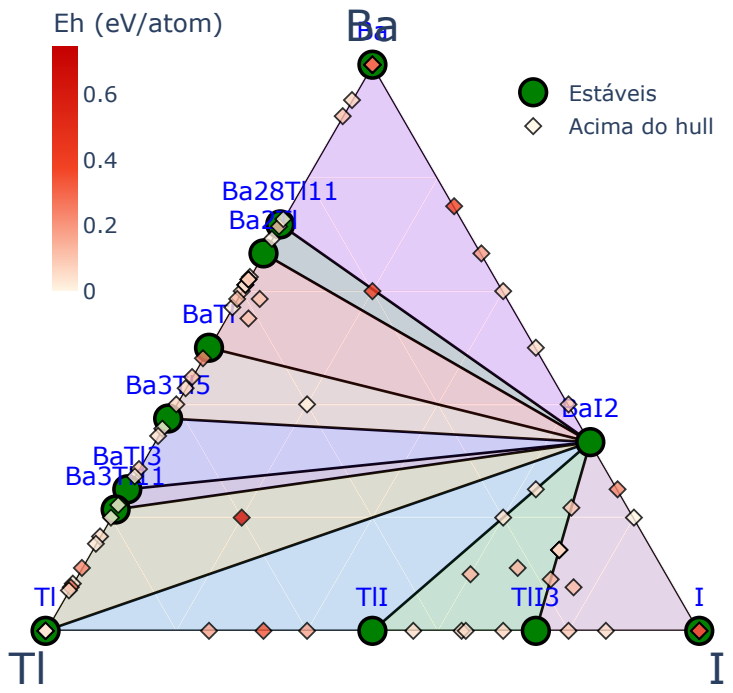
Fonte: elaborado pelo autor.

Figura A19: Diagrama de fase do sistema Ba-Tl-F.



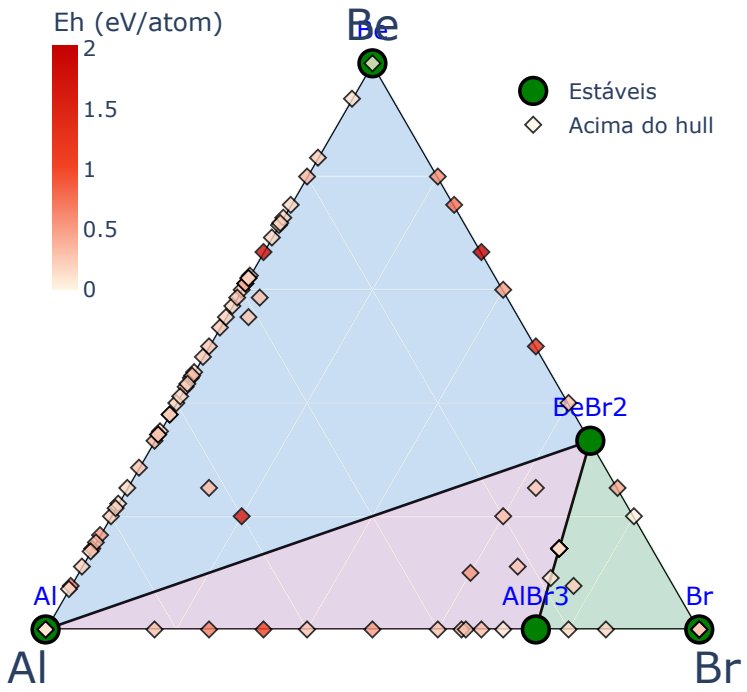
Fonte: elaborado pelo autor.

Figura A20: Diagrama de fase do sistema Ba-Tl-I.



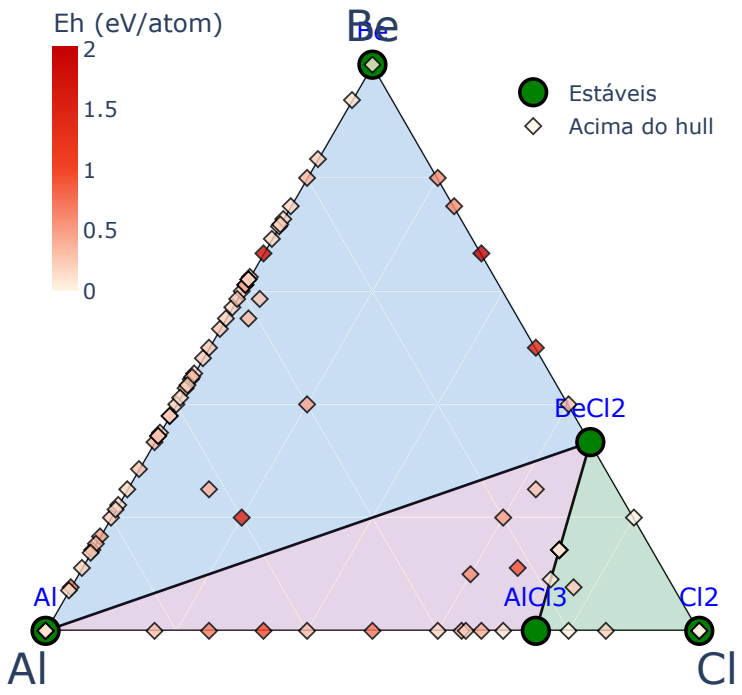
Fonte: elaborado pelo autor.

Figura A21: Diagrama de fase do sistema Be-Al-Br.



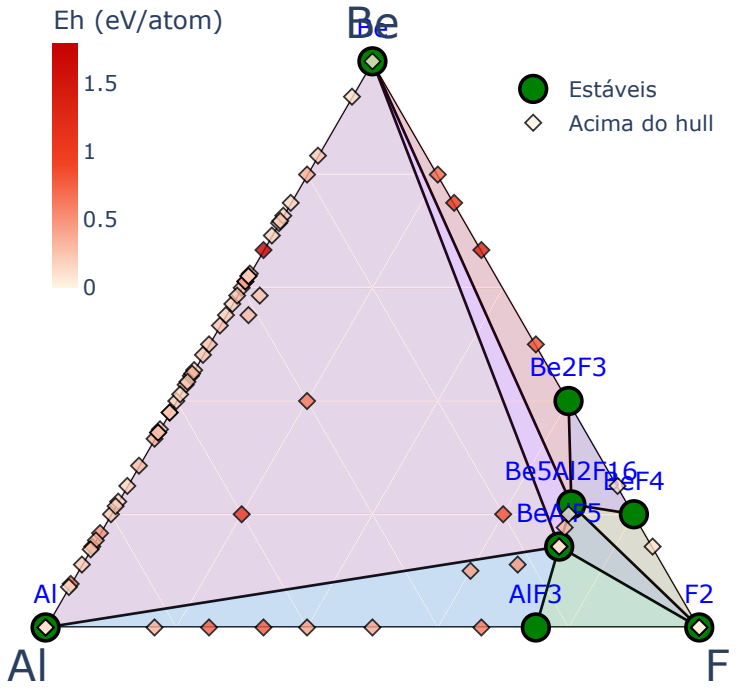
Fonte: elaborado pelo autor.

Figura A22: Diagrama de fase do sistema Be-Al-Cl.



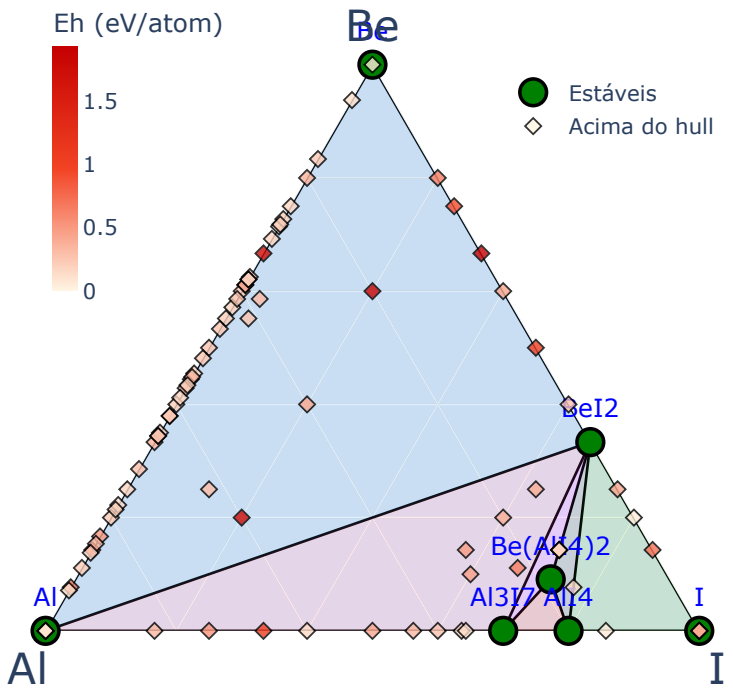
Fonte: elaborado pelo autor.

Figura A23: Diagrama de fase do sistema Be-Al-F.



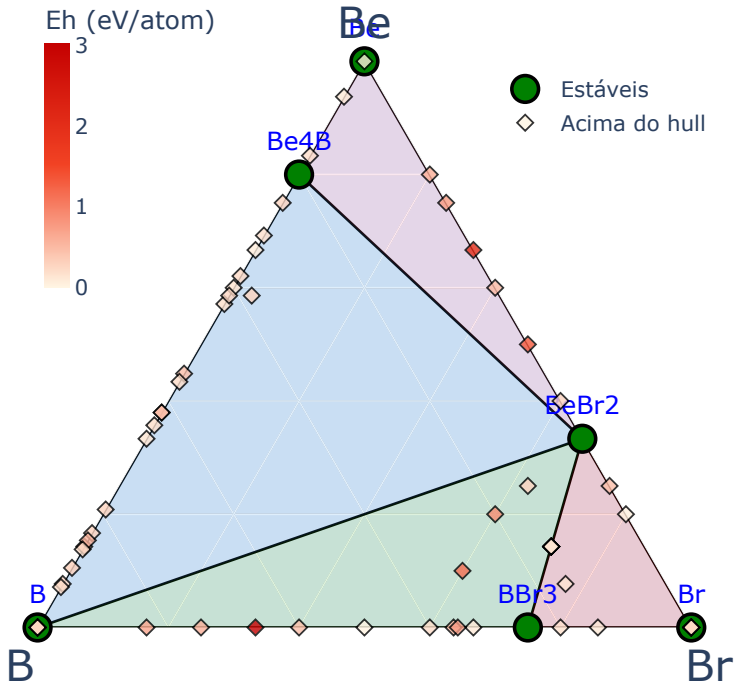
Fonte: elaborado pelo autor.

Figura A24: Diagrama de fase do sistema Be-Al-I.



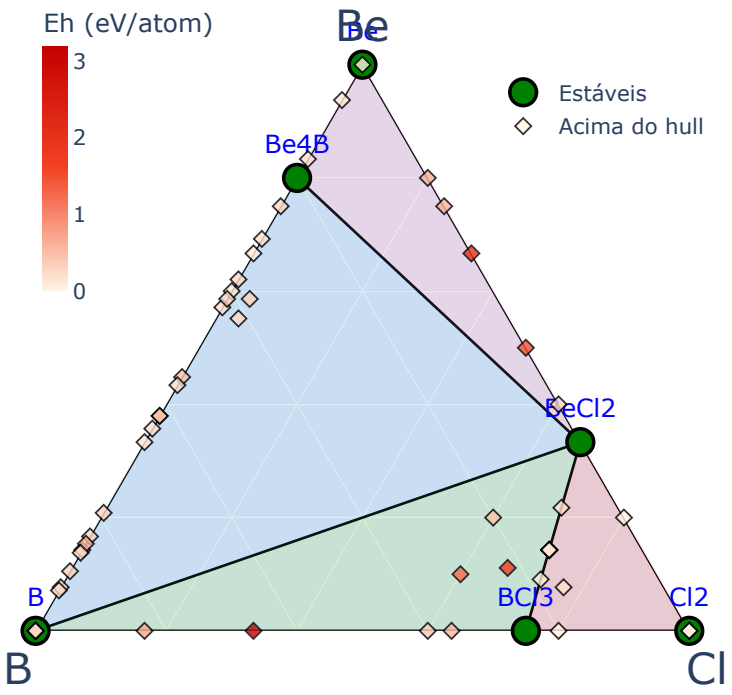
Fonte: elaborado pelo autor.

Figura A25: Diagrama de fase do sistema Be-B-Br.



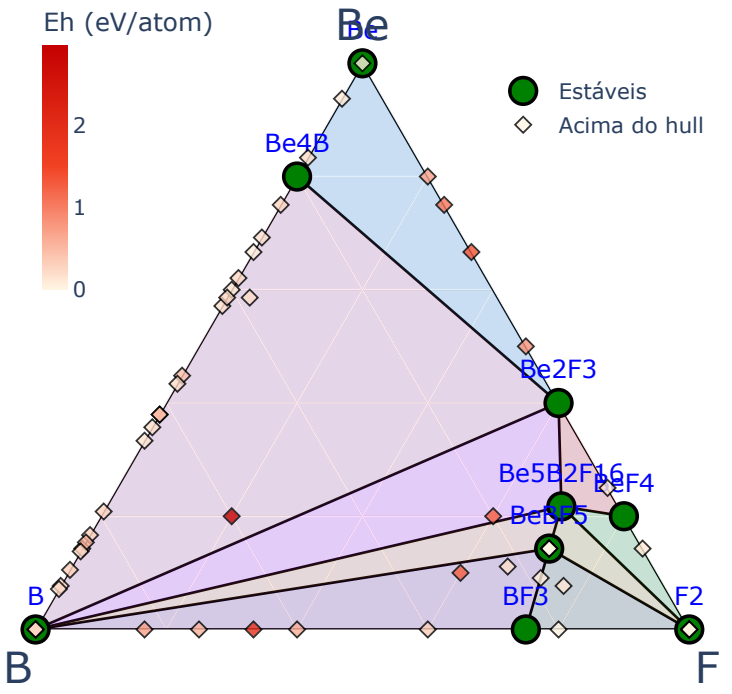
Fonte: elaborado pelo autor.

Figura A26: Diagrama de fase do sistema Be-B-Cl.



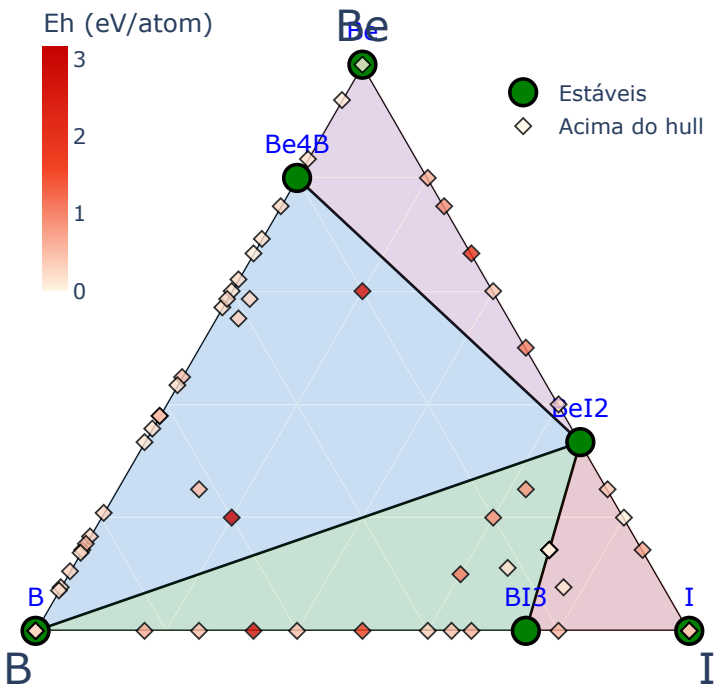
Fonte: elaborado pelo autor.

Figura A27: Diagrama de fase do sistema Be-B-F.



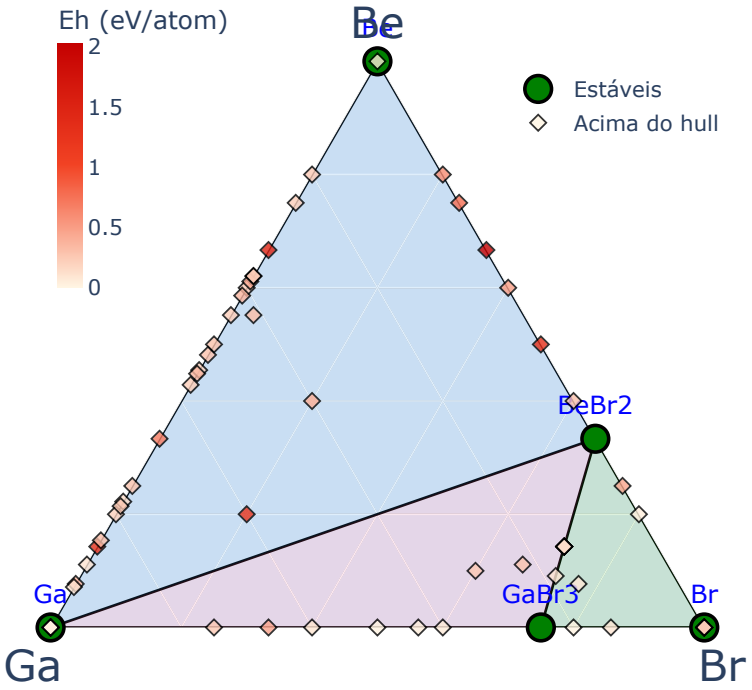
Fonte: elaborado pelo autor.

Figura A28: Diagrama de fase do sistema Be-B-I.



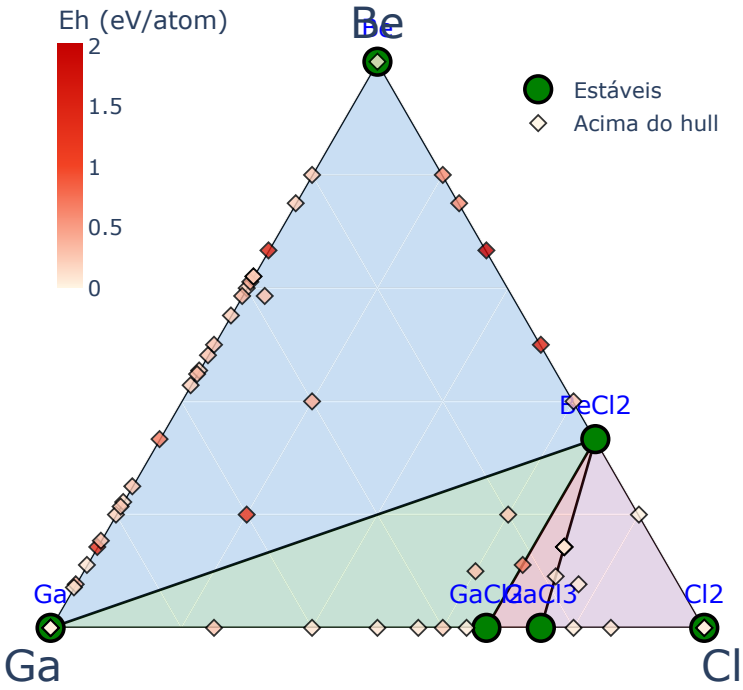
Fonte: elaborado pelo autor.

Figura A29: Diagrama de fase do sistema Be-Ga-Br.



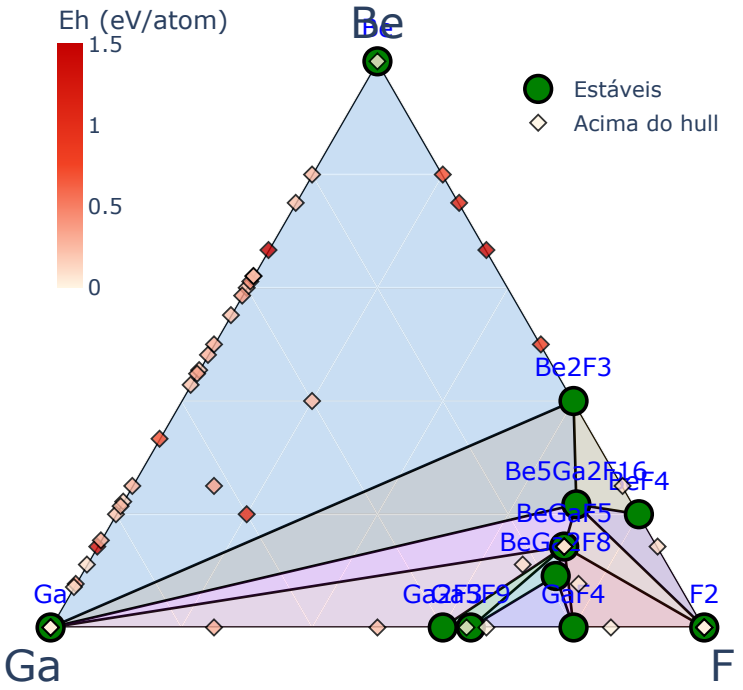
Fonte: elaborado pelo autor.

Figura A30: Diagrama de fase do sistema Be-Ga-Cl.



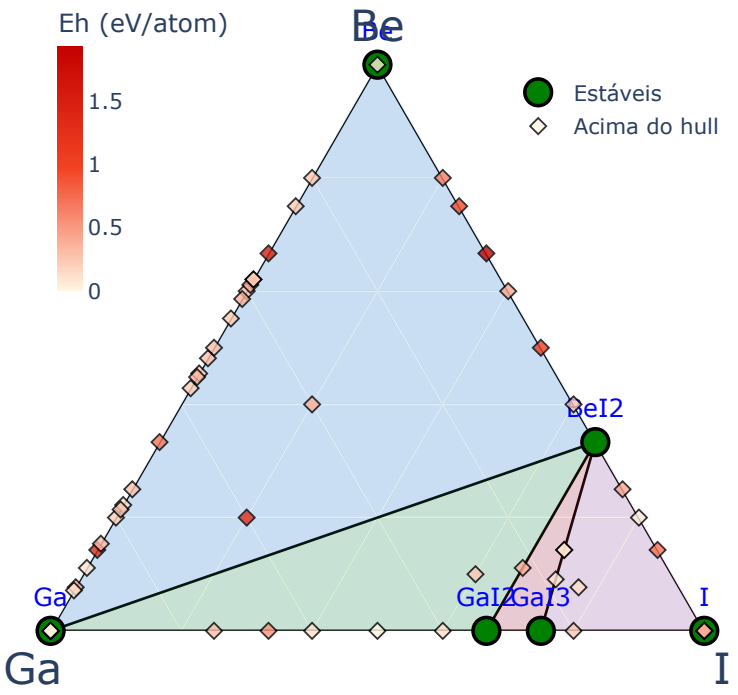
Fonte: elaborado pelo autor.

Figura A31: Diagrama de fase do sistema Be-Ga-F.



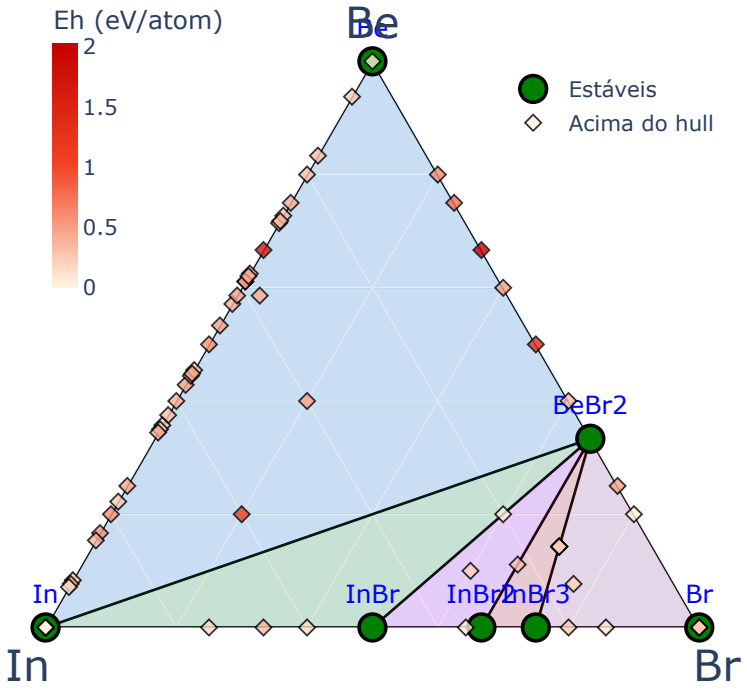
Fonte: elaborado pelo autor.

Figura A32: Diagrama de fase do sistema Be-Ga-I.



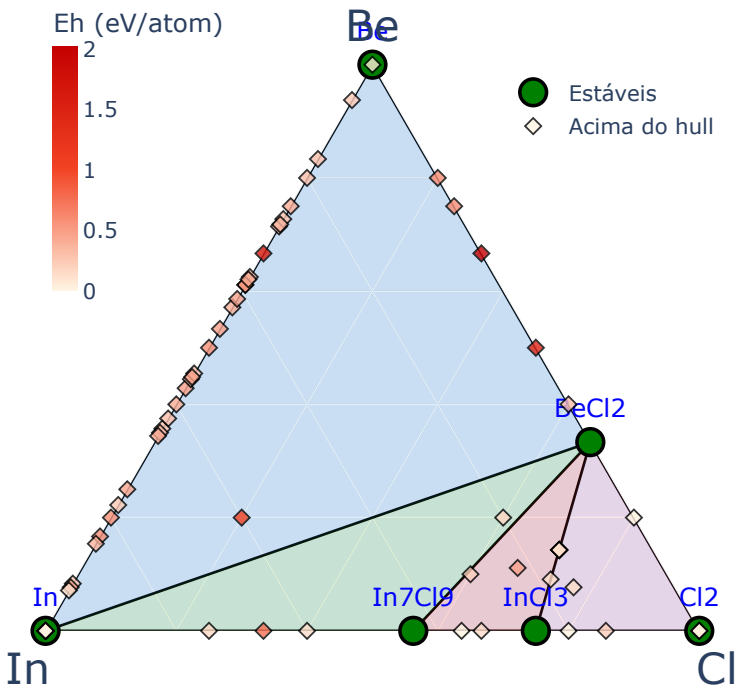
Fonte: elaborado pelo autor.

Figura A33: Diagrama de fase do sistema Be-In-Br.



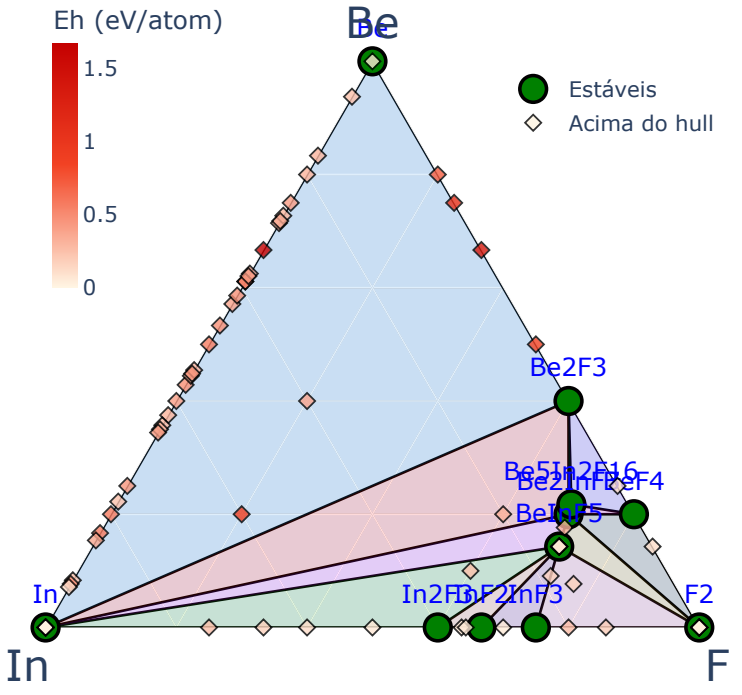
Fonte: elaborado pelo autor.

Figura A34: Diagrama de fase do sistema Be-In-Cl.



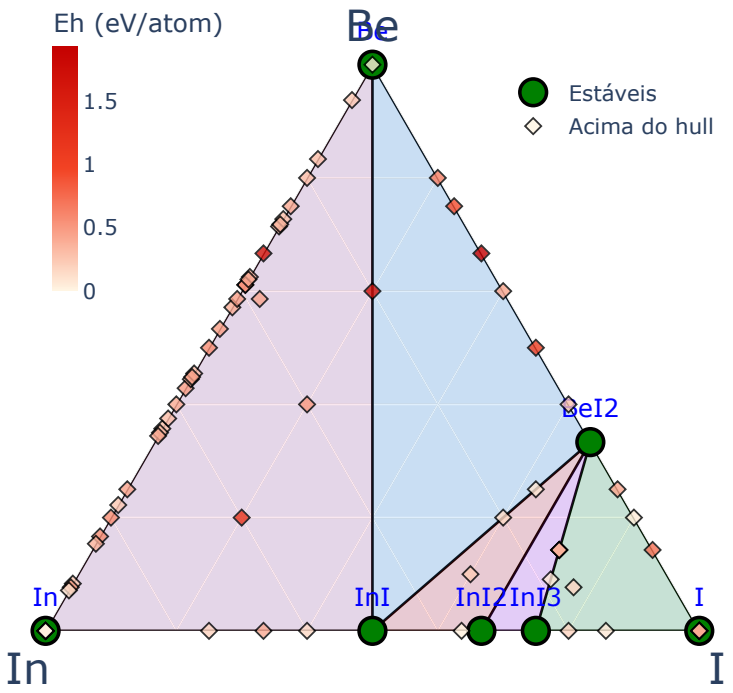
Fonte: elaborado pelo autor.

Figura A35: Diagrama de fase do sistema Be-In-F.



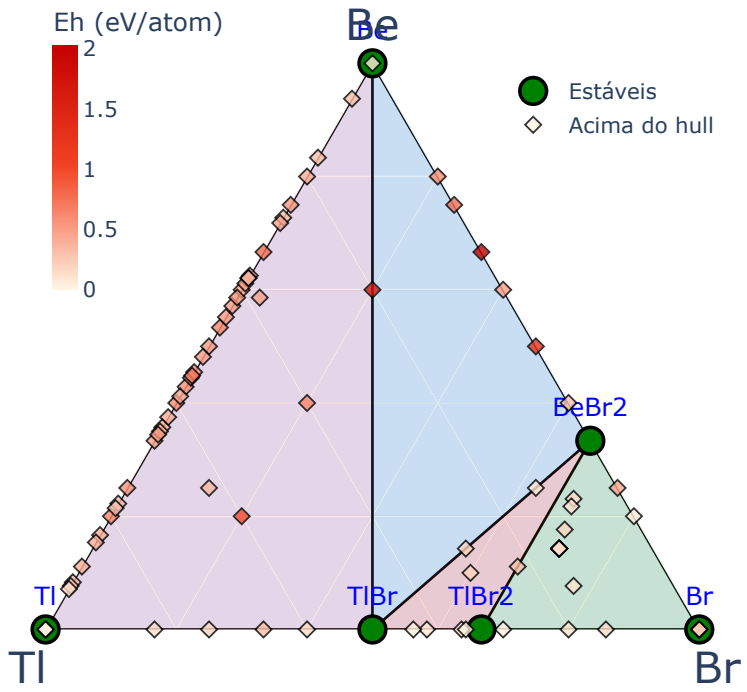
Fonte: elaborado pelo autor.

Figura A36: Diagrama de fase do sistema Be-In-I.



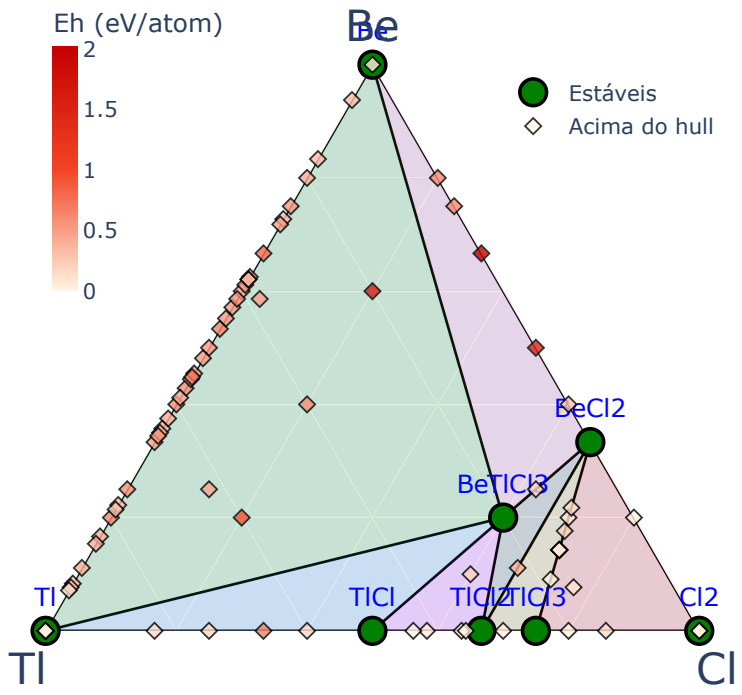
Fonte: elaborado pelo autor.

Figura A37: Diagrama de fase do sistema Be-Tl-Br.



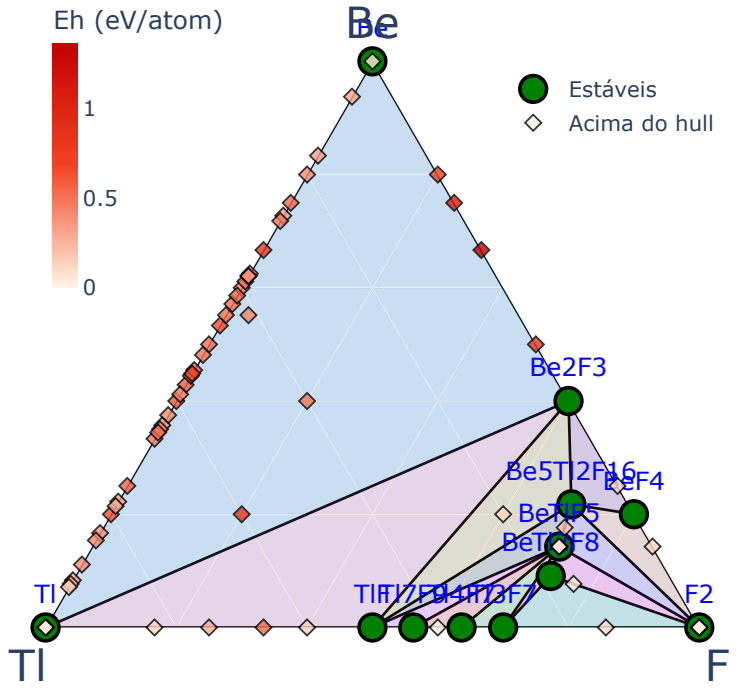
Fonte: elaborado pelo autor.

Figura A38: Diagrama de fase do sistema Be-Tl-Cl.



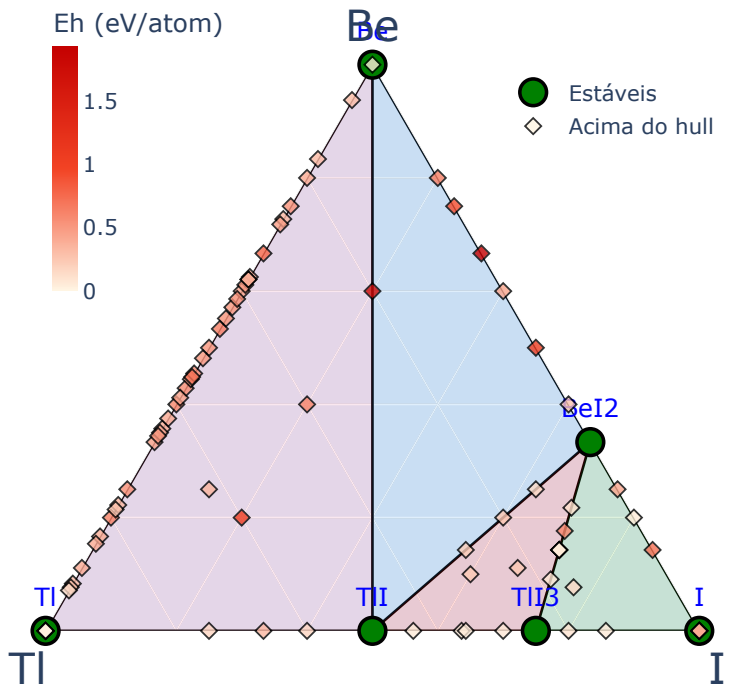
Fonte: elaborado pelo autor.

Figura A39: Diagrama de fase do sistema Be-Tl-F.



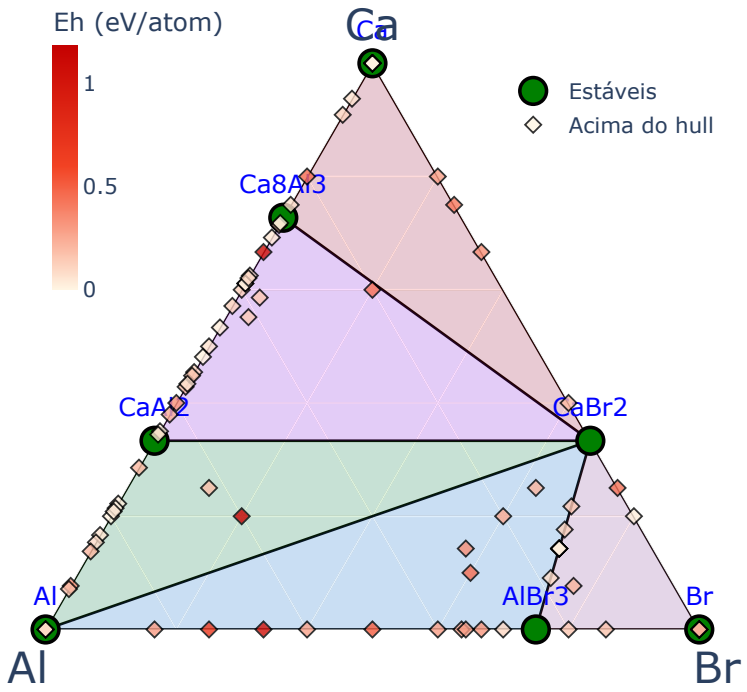
Fonte: elaborado pelo autor.

Figura A40: Diagrama de fase do sistema Be-Tl-I.



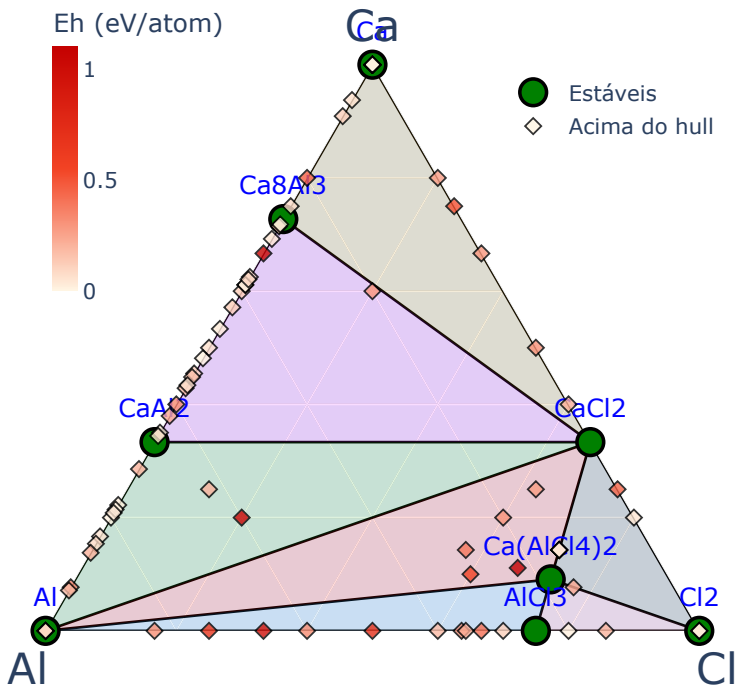
Fonte: elaborado pelo autor.

Figura A41: Diagrama de fase do sistema Ca-Al-Br.



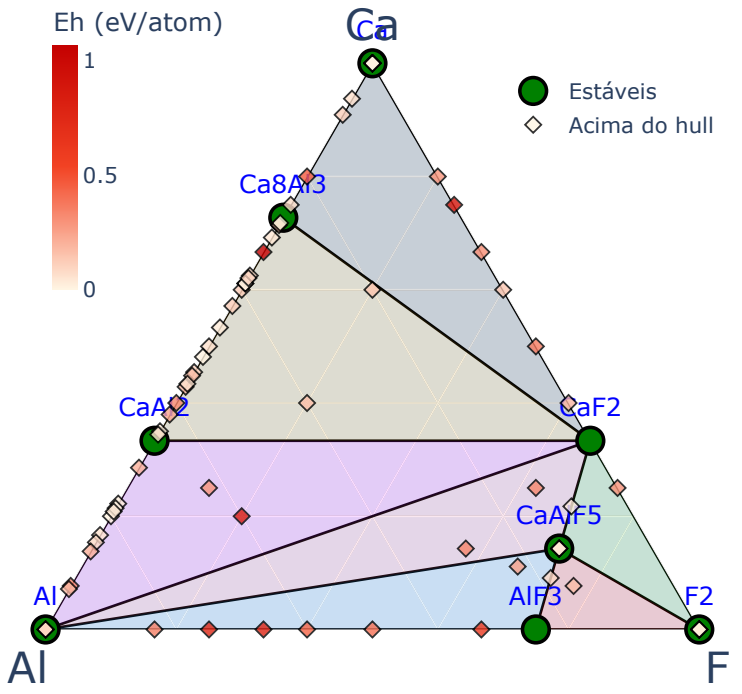
Fonte: elaborado pelo autor.

Figura A42: Diagrama de fase do sistema Ca-Al-Cl.



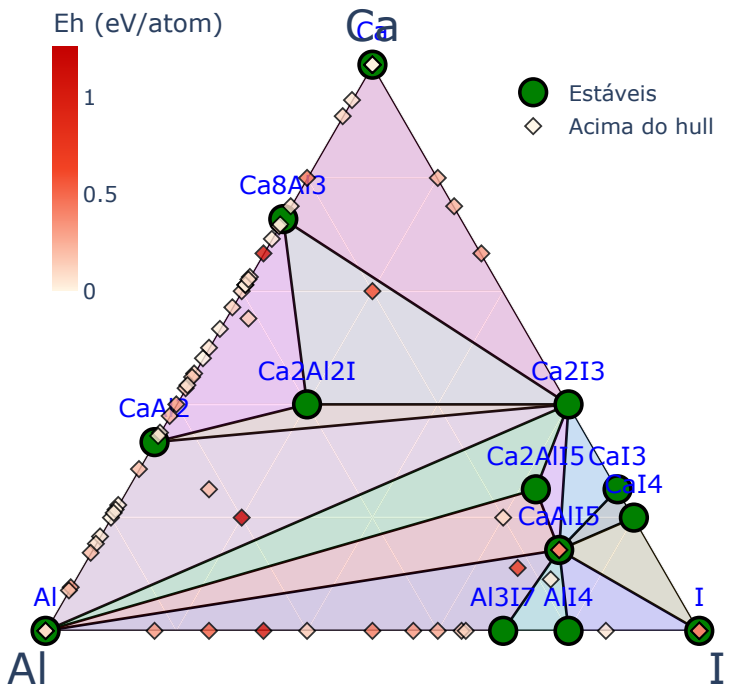
Fonte: elaborado pelo autor.

Figura A43: Diagrama de fase do sistema Ca-Al-F.



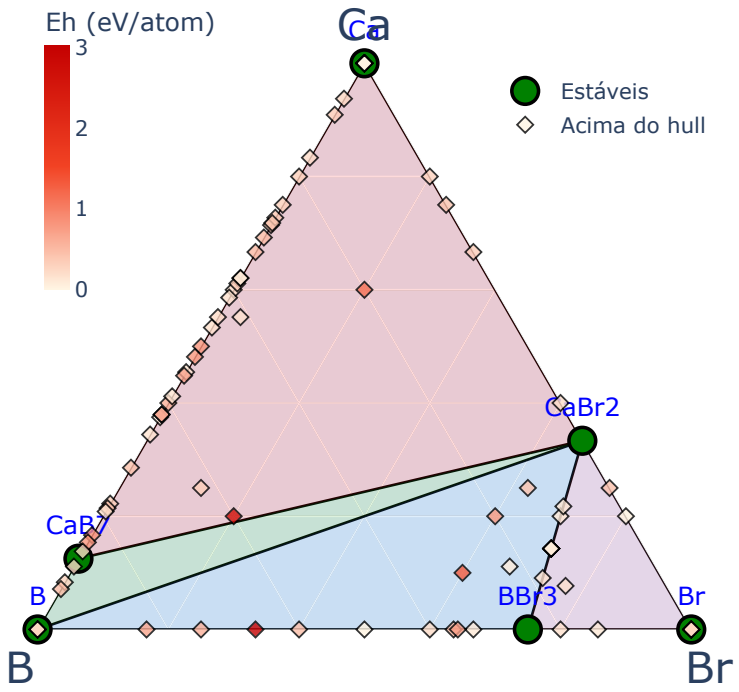
Fonte: elaborado pelo autor.

Figura A44: Diagrama de fase do sistema Ca-Al-I.



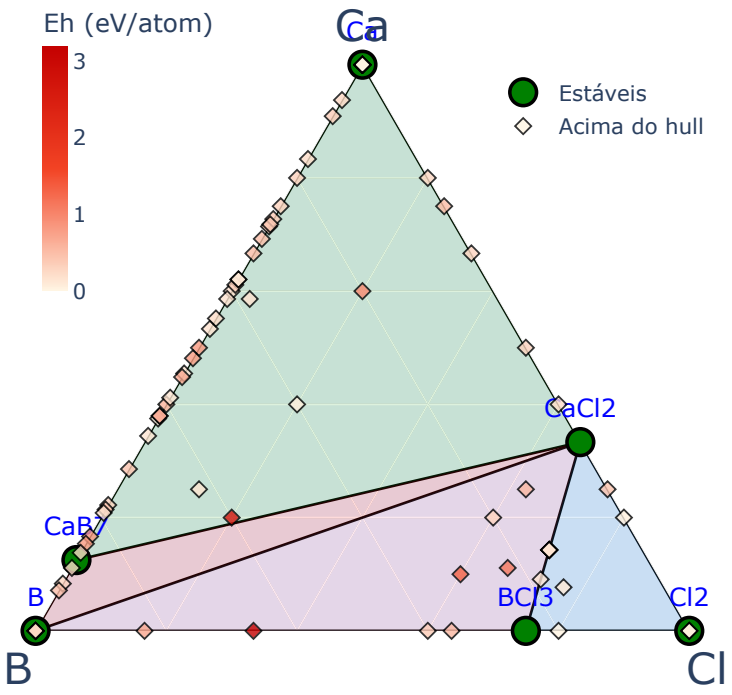
Fonte: elaborado pelo autor.

Figura A45: Diagrama de fase do sistema Ca-B-Br.



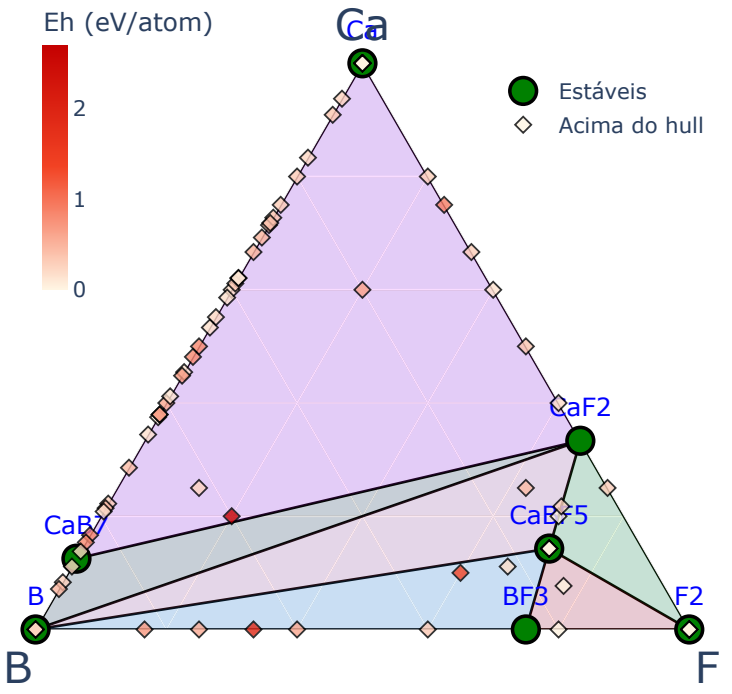
Fonte: elaborado pelo autor.

Figura A46: Diagrama de fase do sistema Ca-B-Cl.



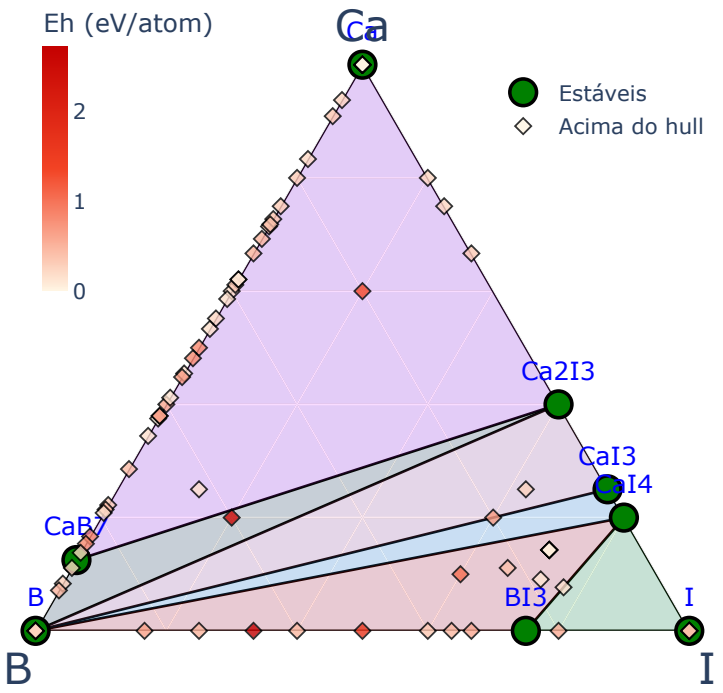
Fonte: elaborado pelo autor.

Figura A47: Diagrama de fase do sistema Ca-B-F.



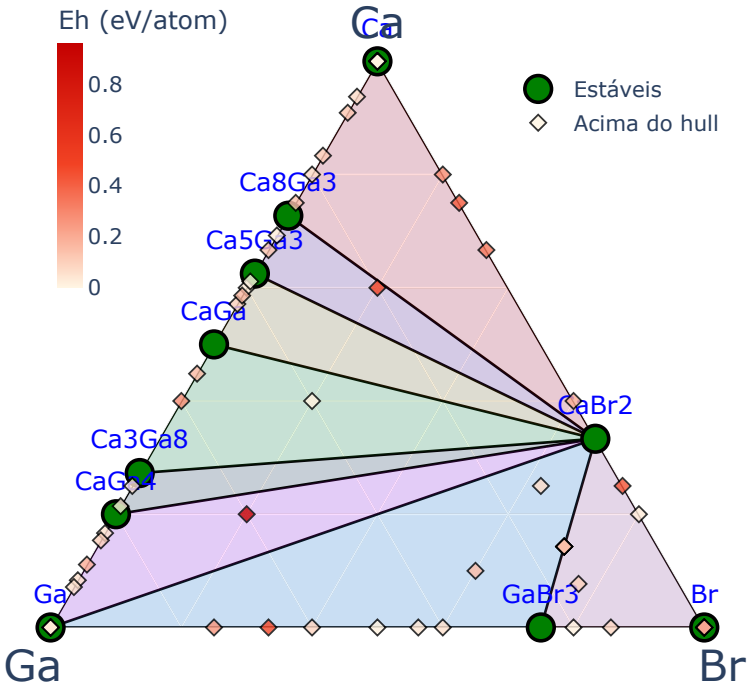
Fonte: elaborado pelo autor.

Figura A48: Diagrama de fase do sistema Ca-B-I.



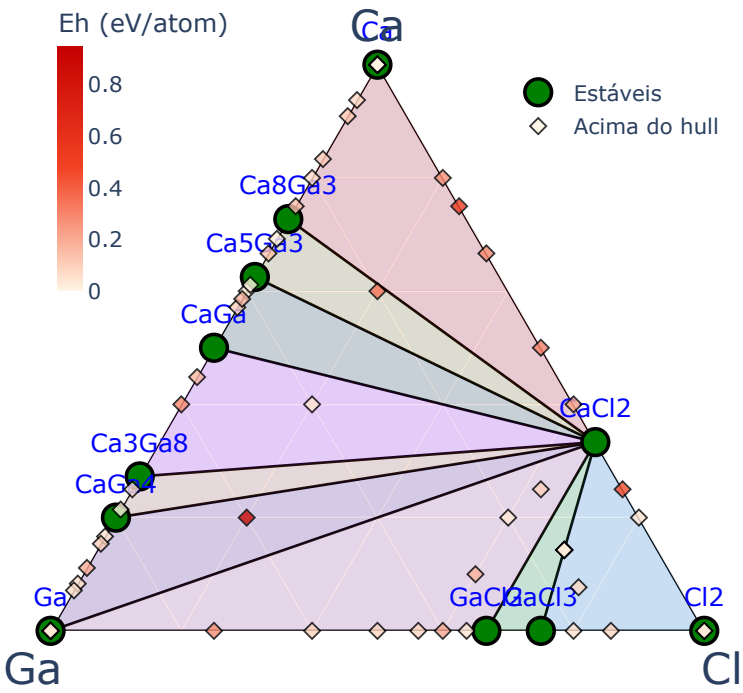
Fonte: elaborado pelo autor.

Figura A49: Diagrama de fase do sistema Ca-Ga-Br.



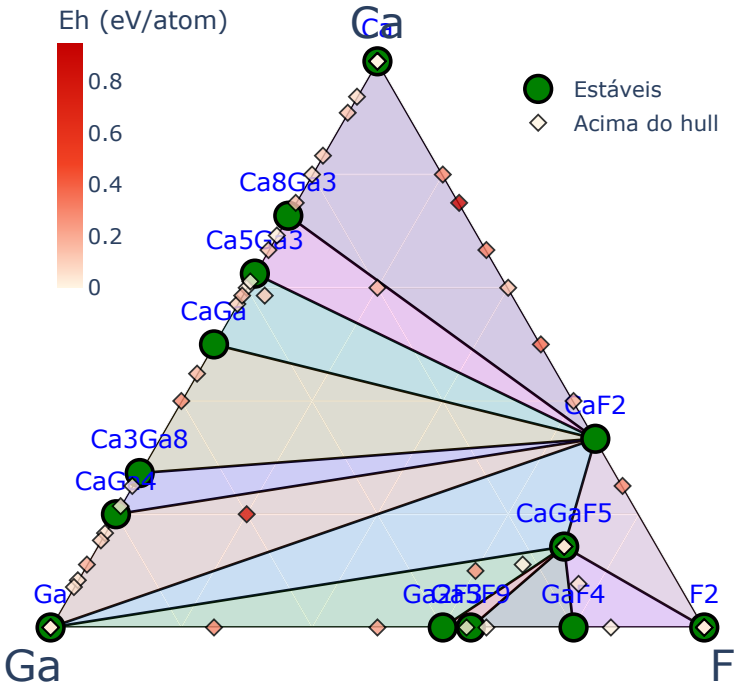
Fonte: elaborado pelo autor.

Figura A50: Diagrama de fase do sistema Ca-Ga-Cl.



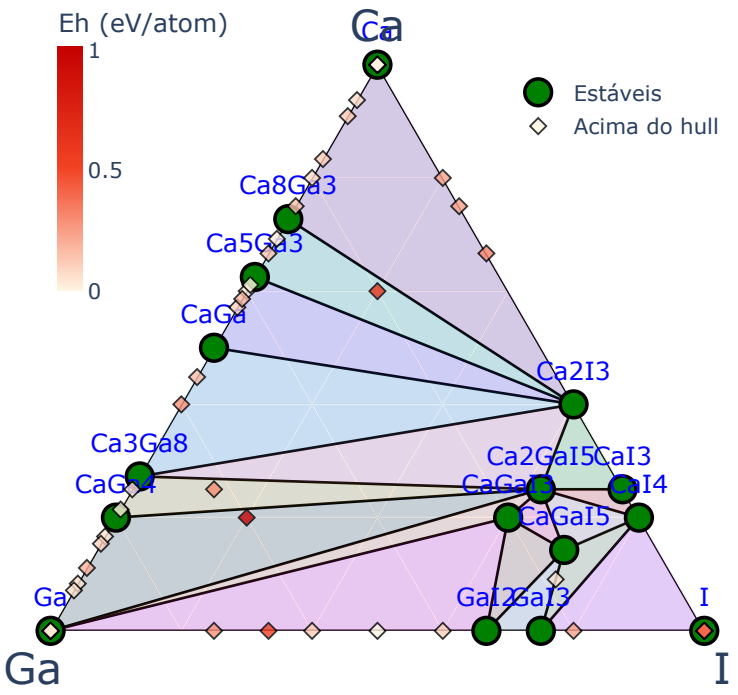
Fonte: elaborado pelo autor.

Figura A51: Diagrama de fase do sistema Ca-Ga-F.



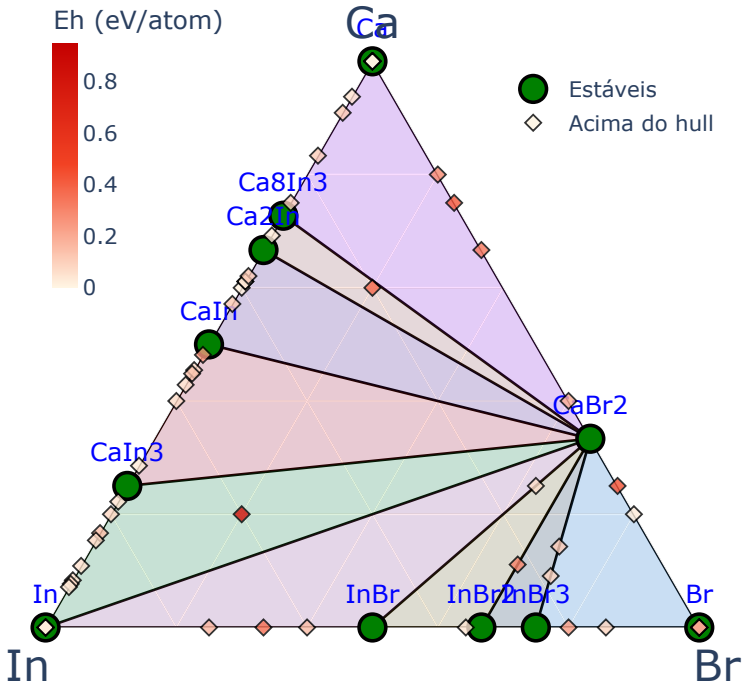
Fonte: elaborado pelo autor.

Figura A52: Diagrama de fase do sistema Ca-Ga-I.



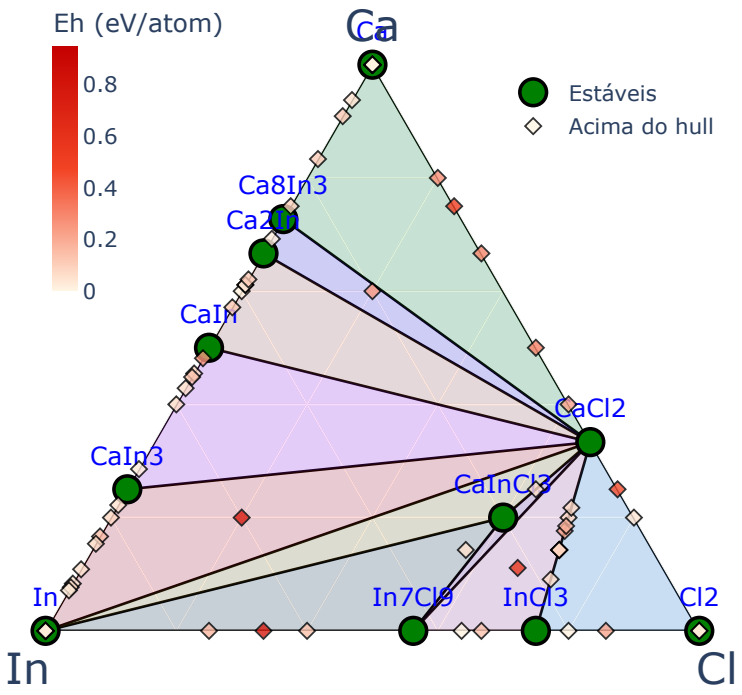
Fonte: elaborado pelo autor.

Figura A53: Diagrama de fase do sistema Ca-In-Br.



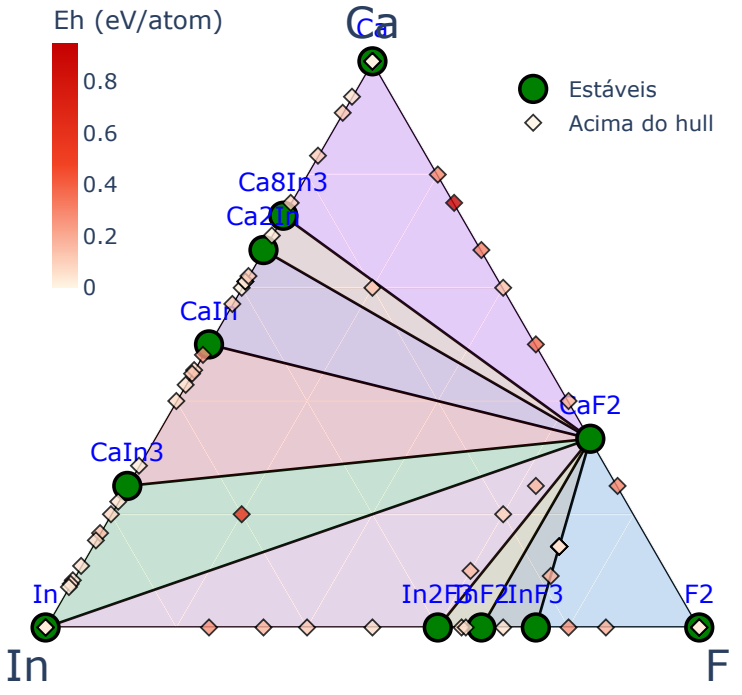
Fonte: elaborado pelo autor.

Figura A54: Diagrama de fase do sistema Ca-In-Cl.



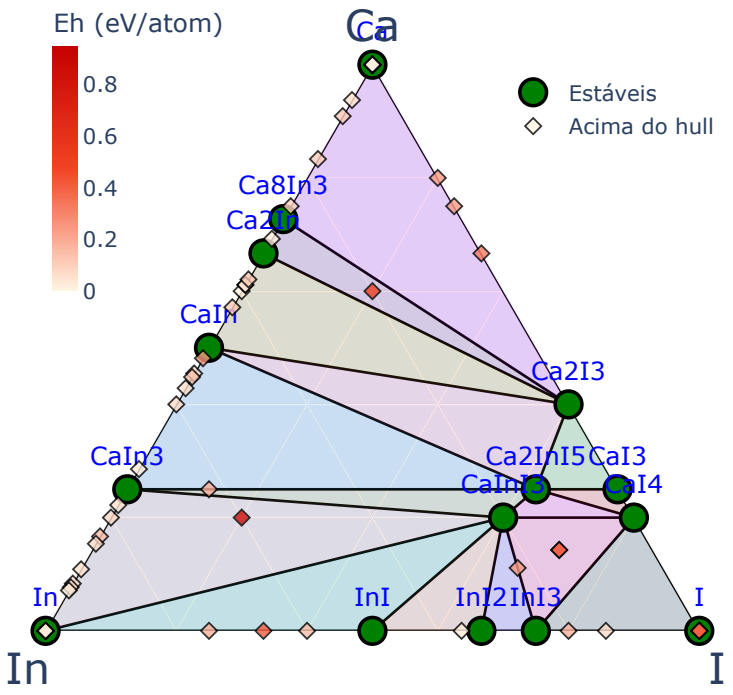
Fonte: elaborado pelo autor.

Figura A55: Diagrama de fase do sistema Ca-In-F.



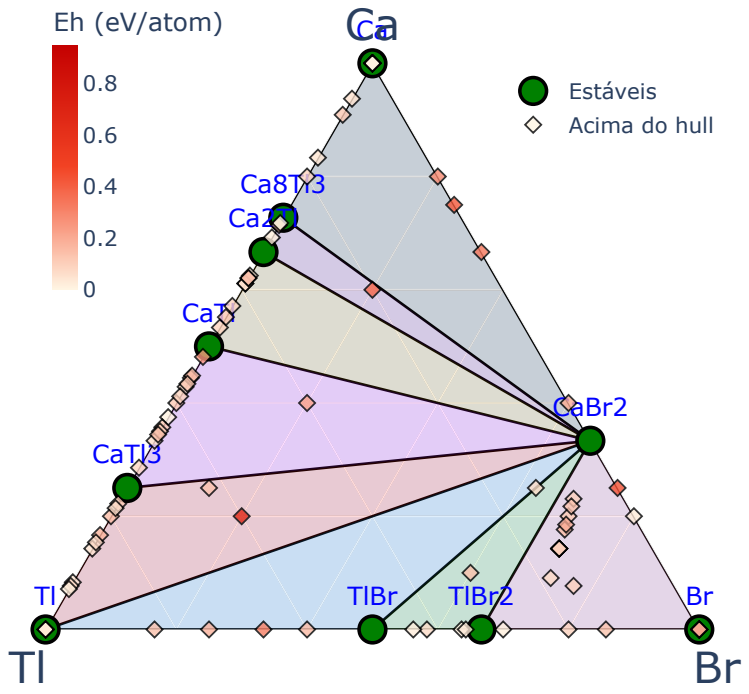
Fonte: elaborado pelo autor.

Figura A56: Diagrama de fase do sistema Ca-In-I.



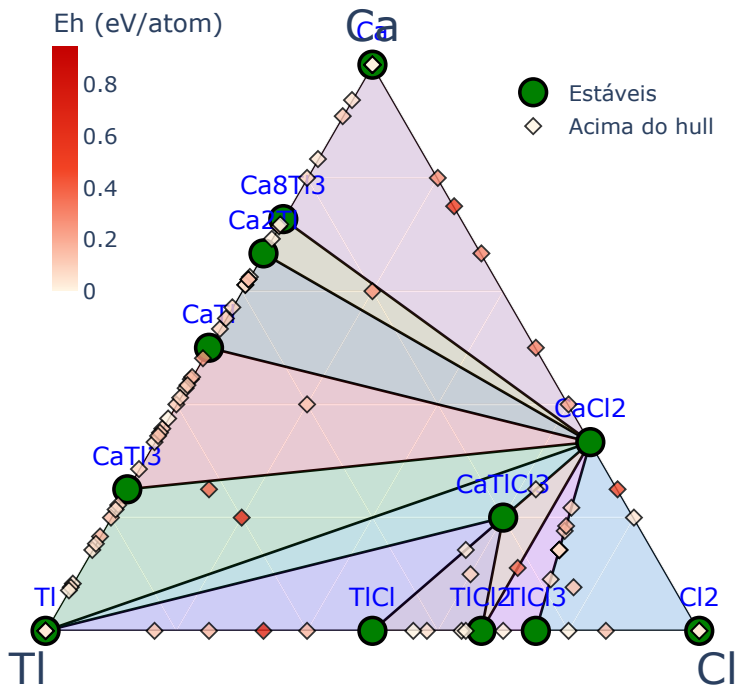
Fonte: elaborado pelo autor.

Figura A57: Diagrama de fase do sistema Ca-Tl-Br.



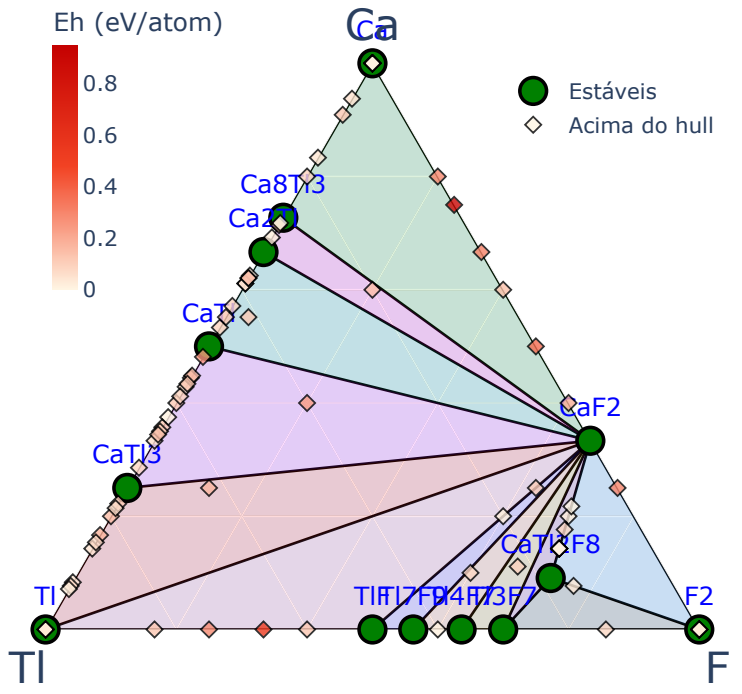
Fonte: elaborado pelo autor.

Figura A58: Diagrama de fase do sistema Ca-Tl-Cl.



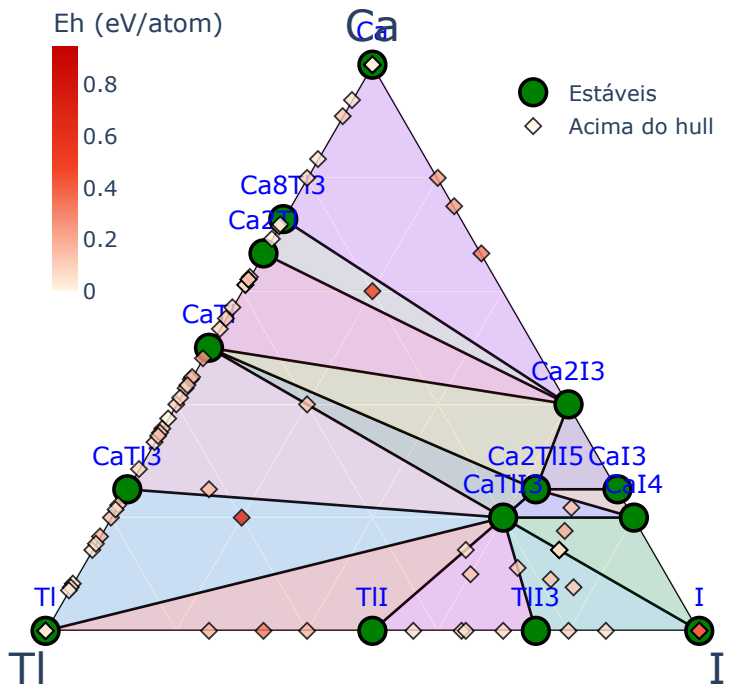
Fonte: elaborado pelo autor.

Figura A59: Diagrama de fase do sistema Ca-Tl-F.



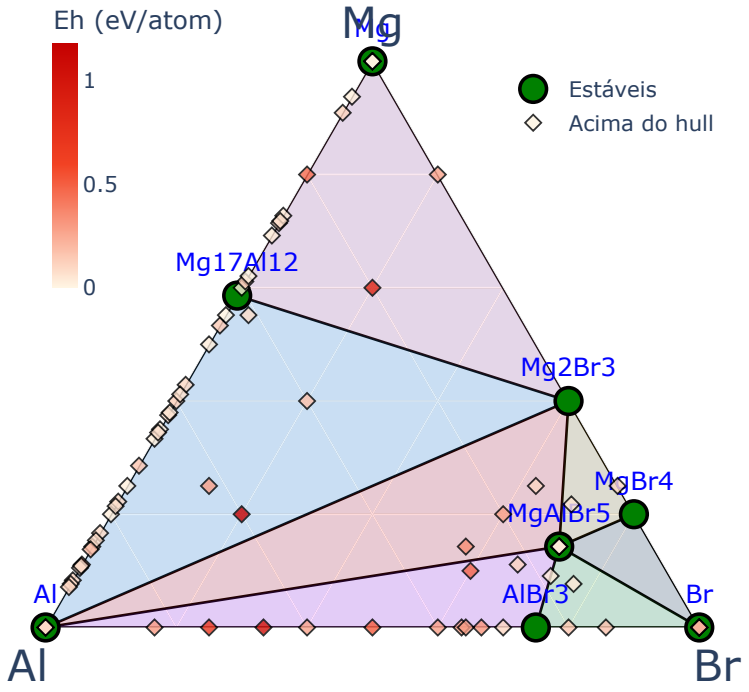
Fonte: elaborado pelo autor.

Figura A60: Diagrama de fase do sistema Ca-Tl-I.



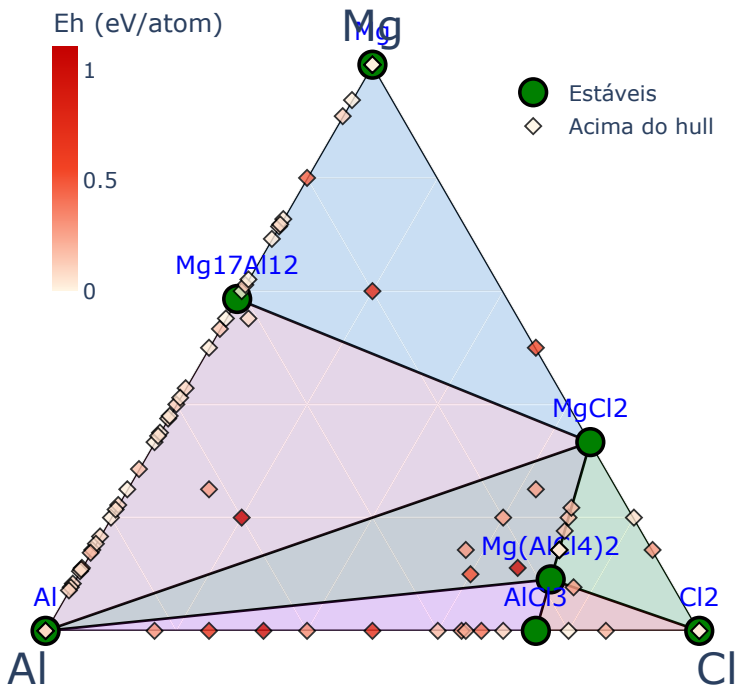
Fonte: elaborado pelo autor.

Figura A61: Diagrama de fase do sistema Mg-Al-Br.



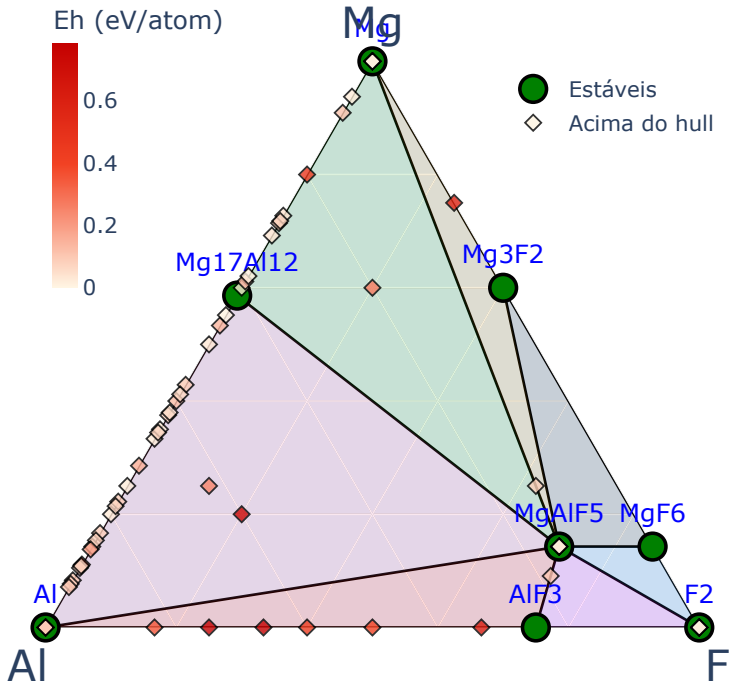
Fonte: elaborado pelo autor.

Figura A62: Diagrama de fase do sistema Mg-Al-Cl.



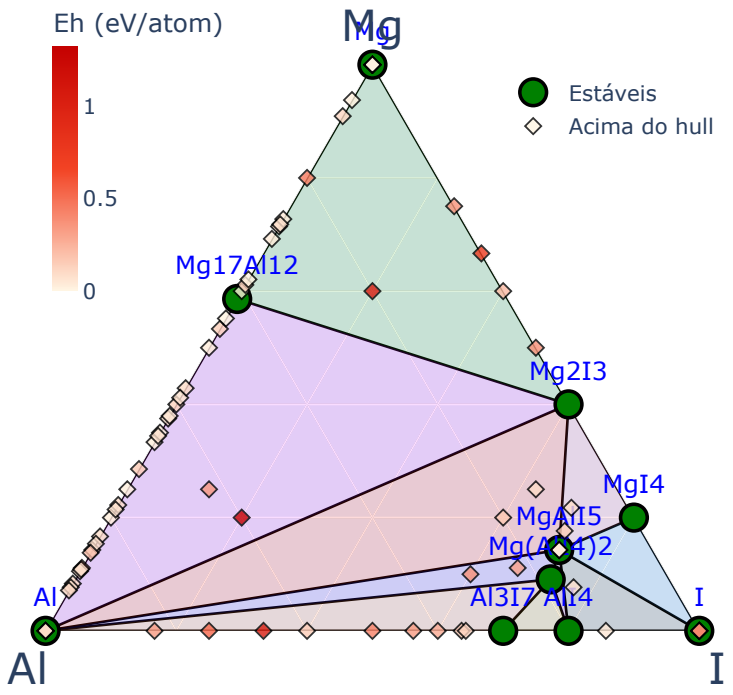
Fonte: elaborado pelo autor.

Figura A63: Diagrama de fase do sistema Mg-Al-F.



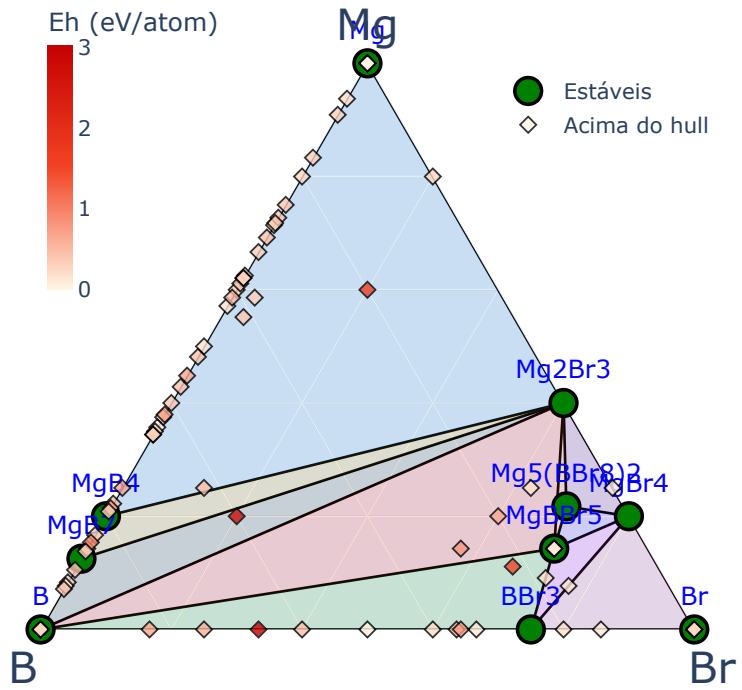
Fonte: elaborado pelo autor.

Figura A64: Diagrama de fase do sistema Mg-Al-I.



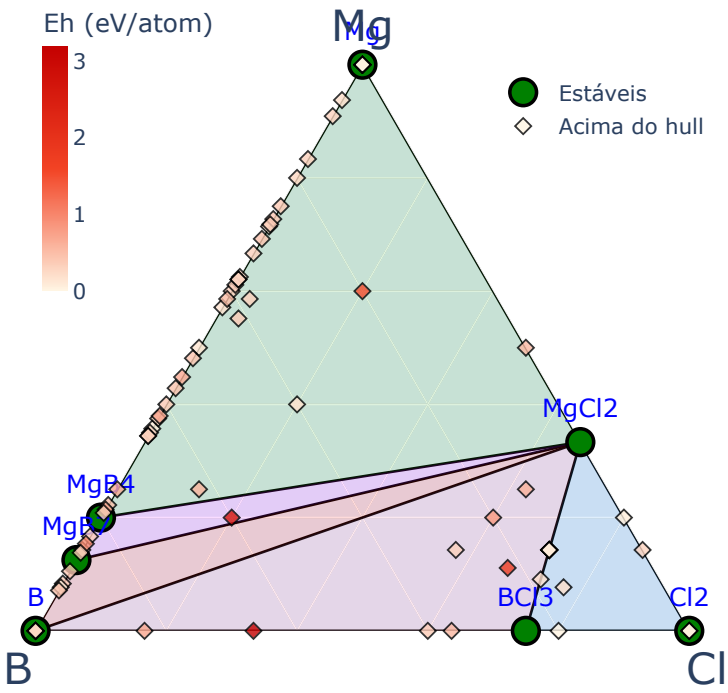
Fonte: elaborado pelo autor.

Figura A65: Diagrama de fase do sistema Mg-B-Br.



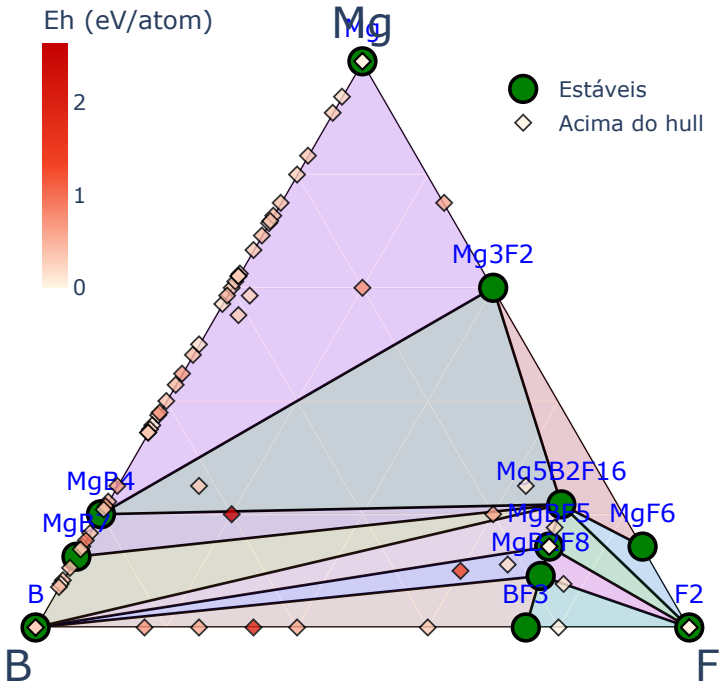
Fonte: elaborado pelo autor.

Figura A66: Diagrama de fase do sistema Mg-B-Cl.



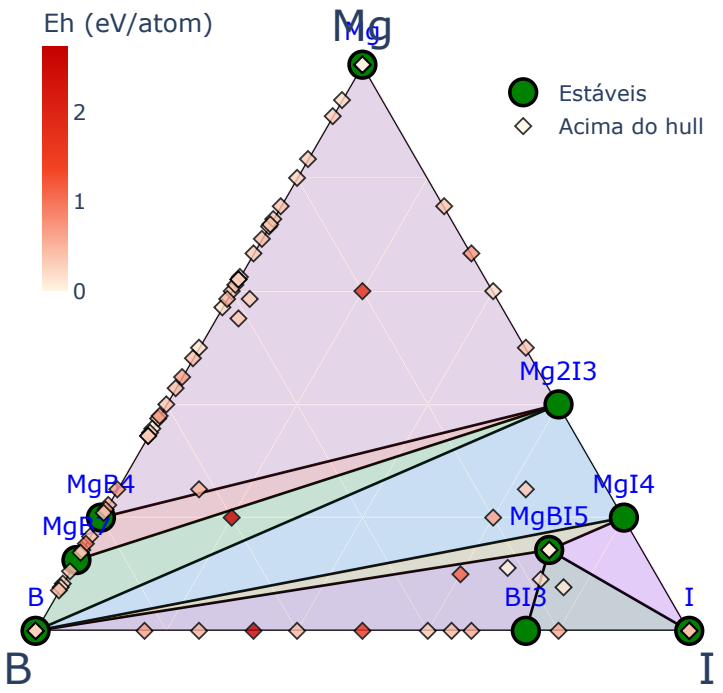
Fonte: elaborado pelo autor.

Figura A67: Diagrama de fase do sistema Mg-B-F.



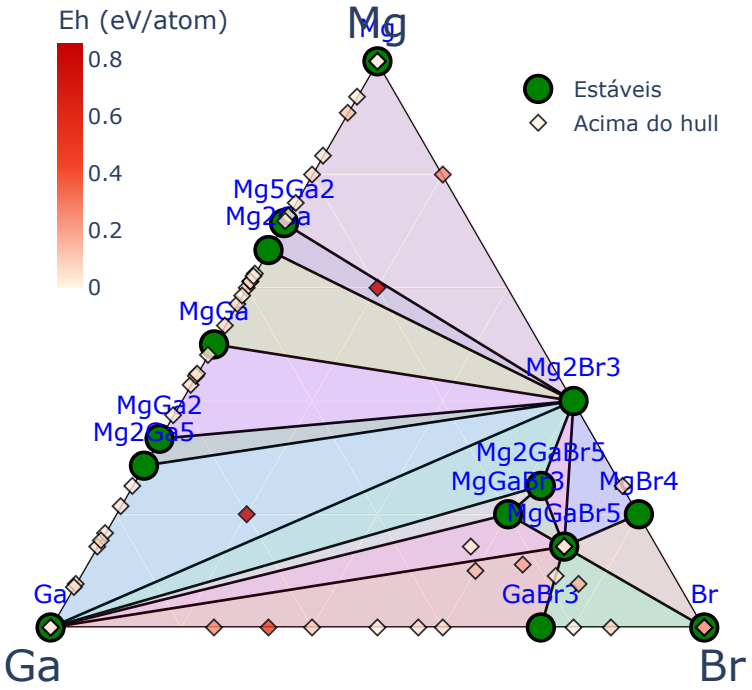
Fonte: elaborado pelo autor.

Figura A68: Diagrama de fase do sistema Mg-B-I.



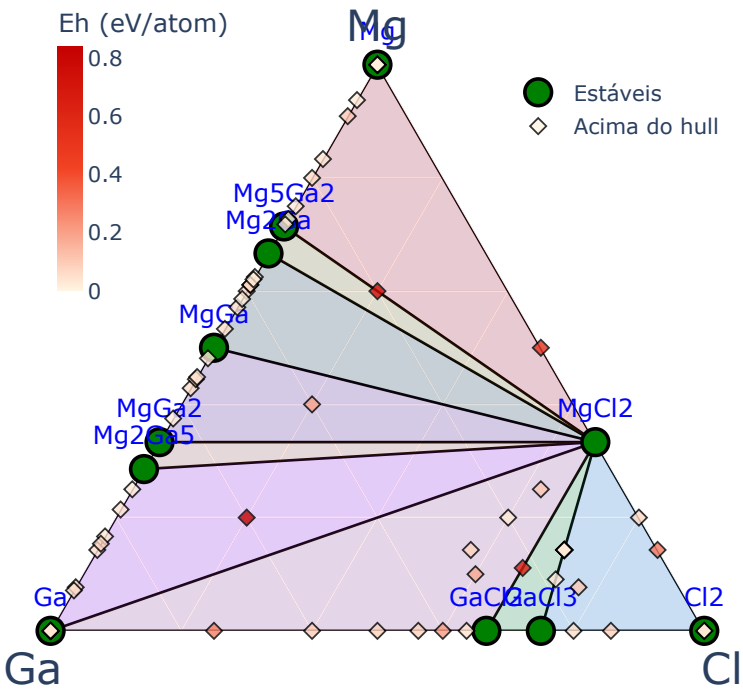
Fonte: elaborado pelo autor.

Figura A69: Diagrama de fase do sistema Mg-Ga-Br.



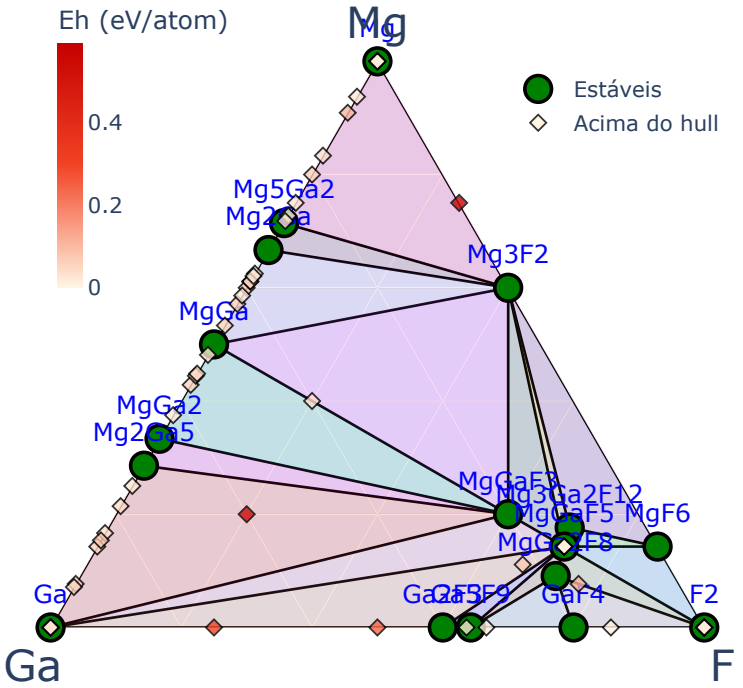
Fonte: elaborado pelo autor.

Figura A70: Diagrama de fase do sistema Mg-Ga-Cl.



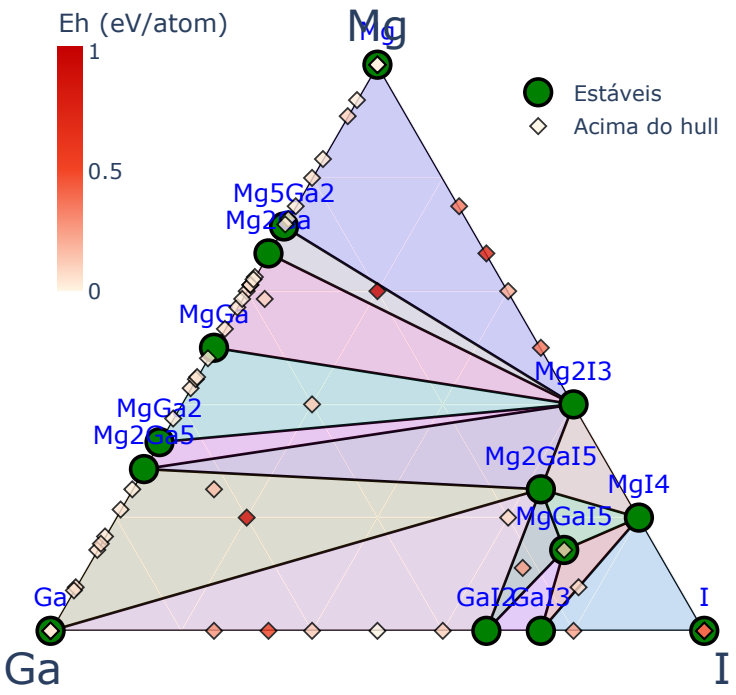
Fonte: elaborado pelo autor.

Figura A71: Diagrama de fase do sistema Mg-Ga-F.



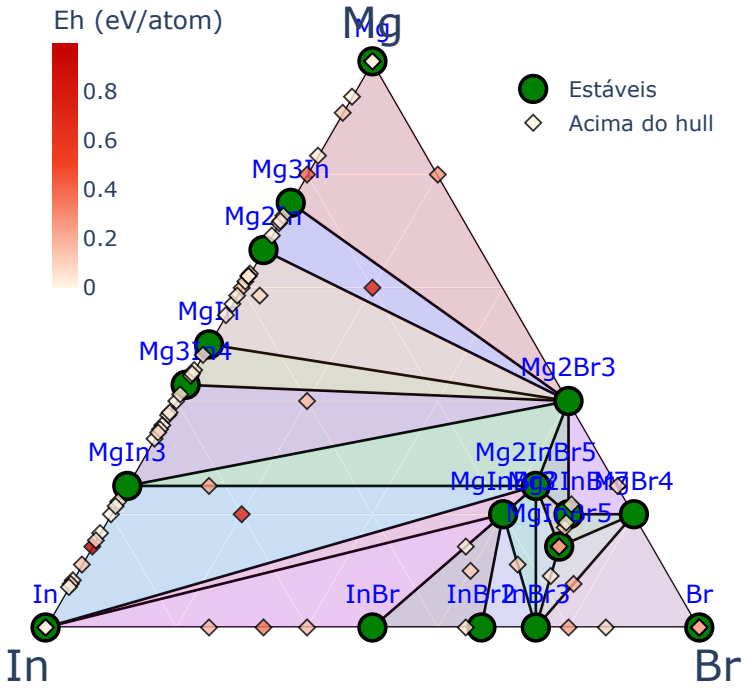
Fonte: elaborado pelo autor.

Figura A72: Diagrama de fase do sistema Mg-Ga-I.



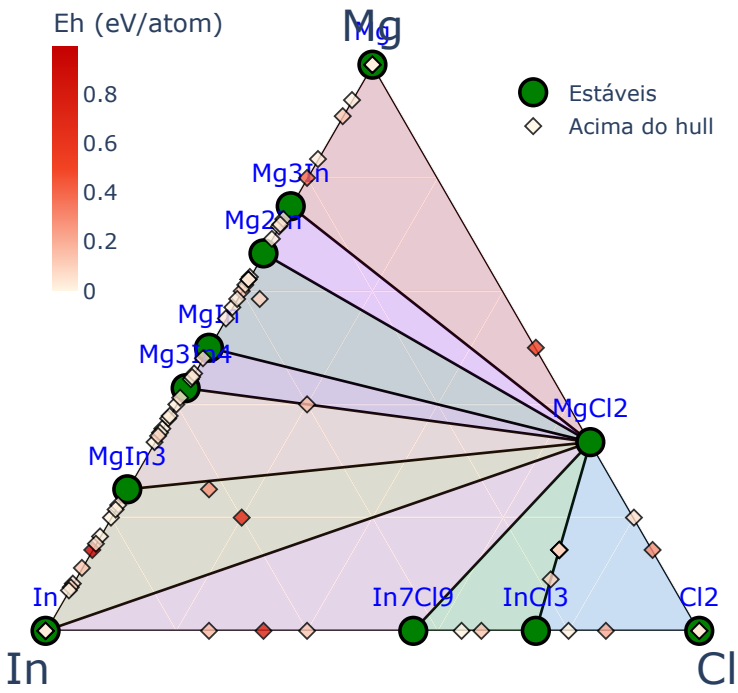
Fonte: elaborado pelo autor.

Figura A73: Diagrama de fase do sistema Mg-In-Br.



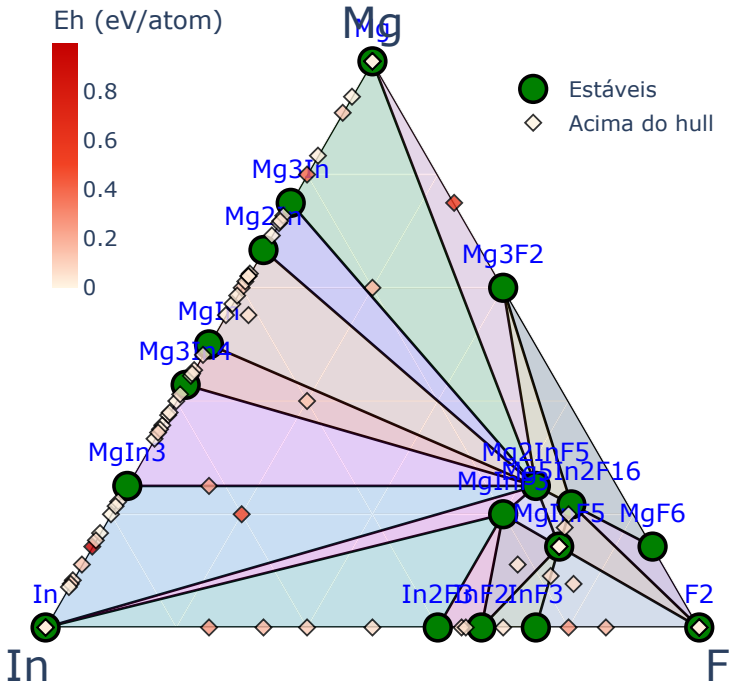
Fonte: elaborado pelo autor.

Figura A74: Diagrama de fase do sistema Mg-In-Cl.



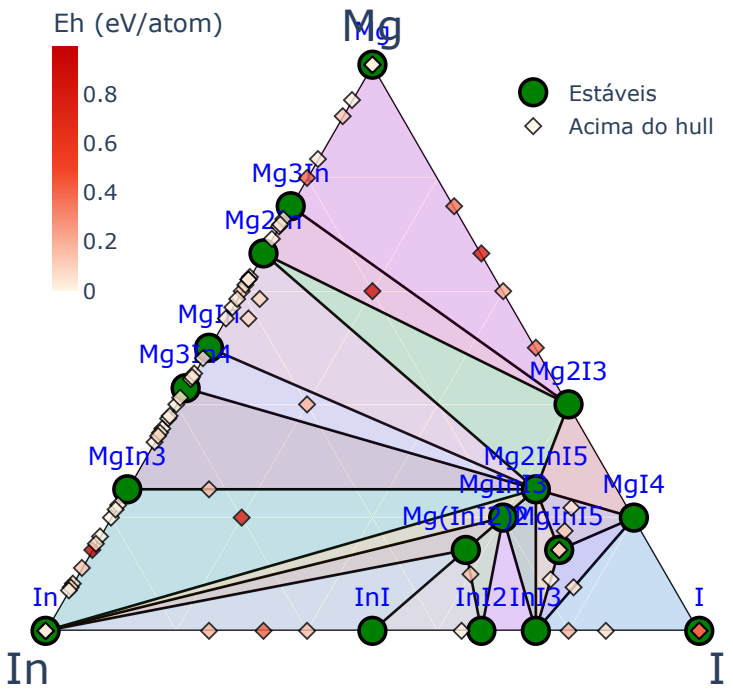
Fonte: elaborado pelo autor.

Figura A75: Diagrama de fase do sistema Mg-In-F.



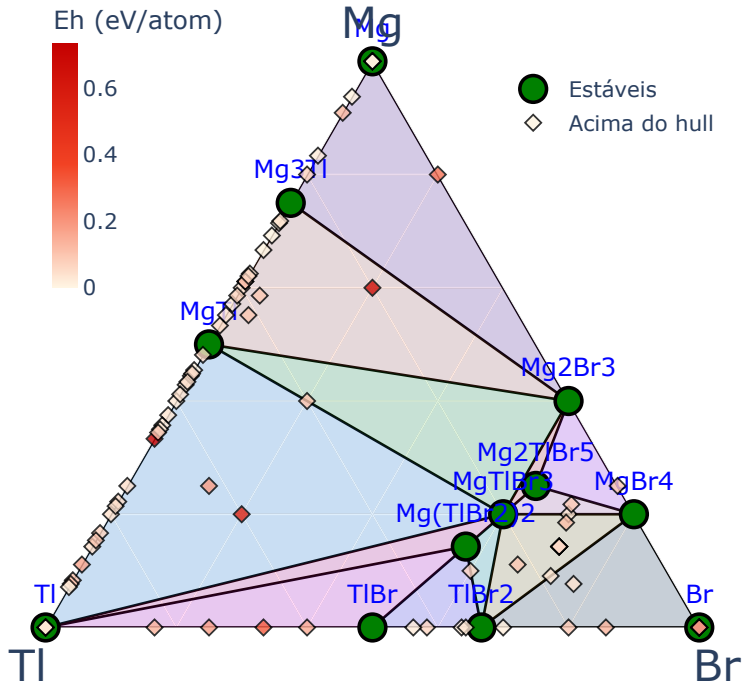
Fonte: elaborado pelo autor.

Figura A76: Diagrama de fase do sistema Mg-In-I.



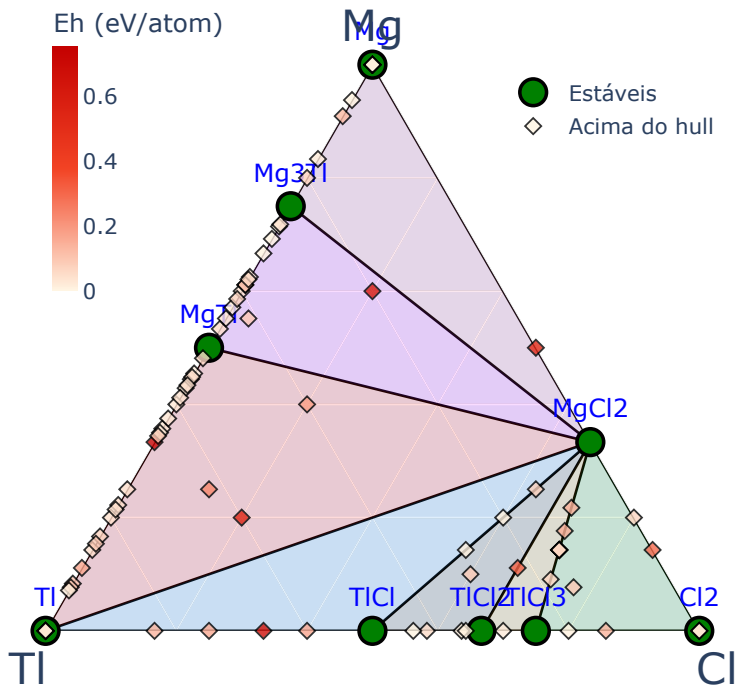
Fonte: elaborado pelo autor.

Figura A77: Diagrama de fase do sistema Mg-Tl-Br.



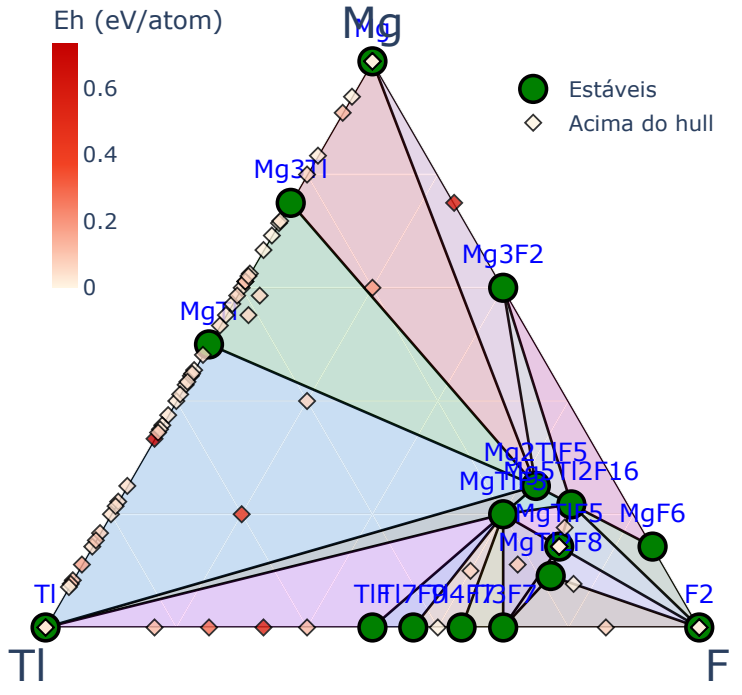
Fonte: elaborado pelo autor.

Figura A78: Diagrama de fase do sistema Mg-Tl-Cl.



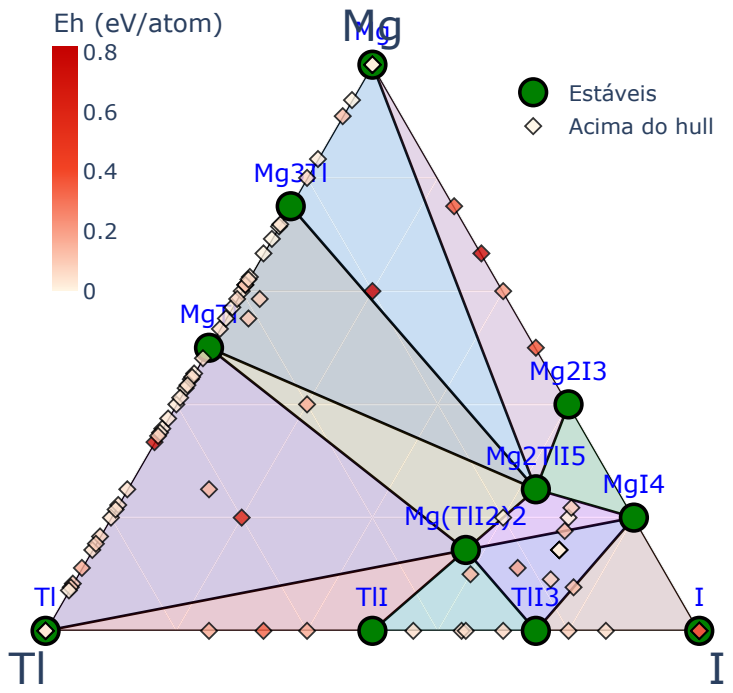
Fonte: elaborado pelo autor.

Figura A79: Diagrama de fase do sistema Mg-Tl-F.



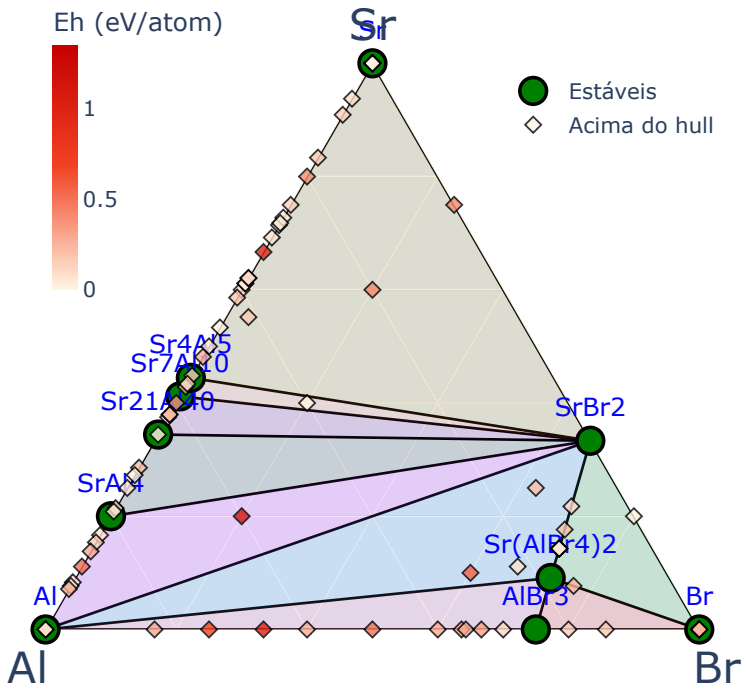
Fonte: elaborado pelo autor.

Figura A80: Diagrama de fase do sistema Mg-Tl-I.



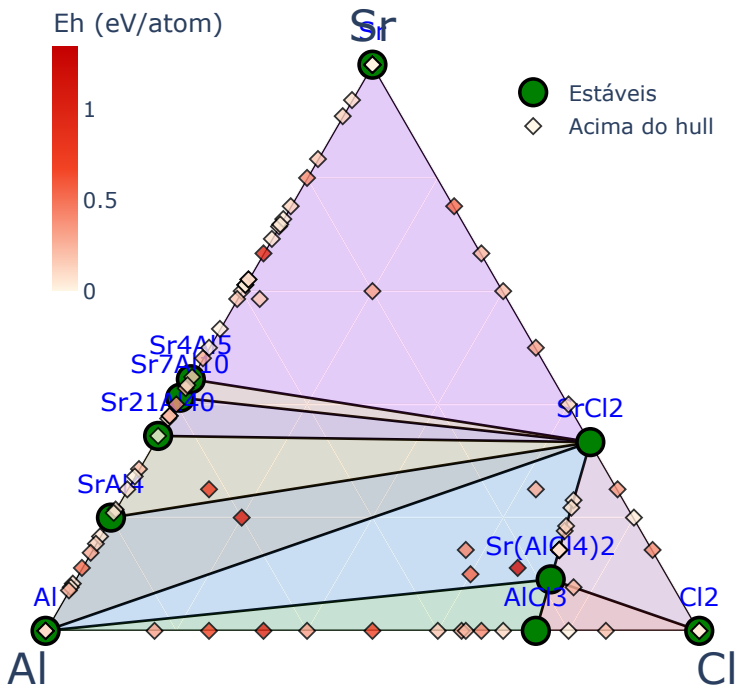
Fonte: elaborado pelo autor.

Figura A81: Diagrama de fase do sistema Sr-Al-Br.



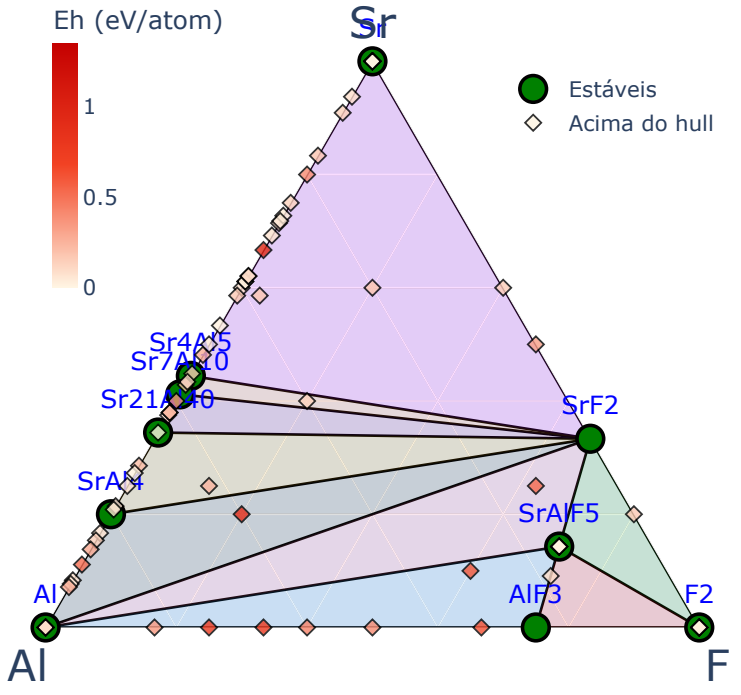
Fonte: elaborado pelo autor.

Figura A82: Diagrama de fase do sistema Sr-Al-Cl.



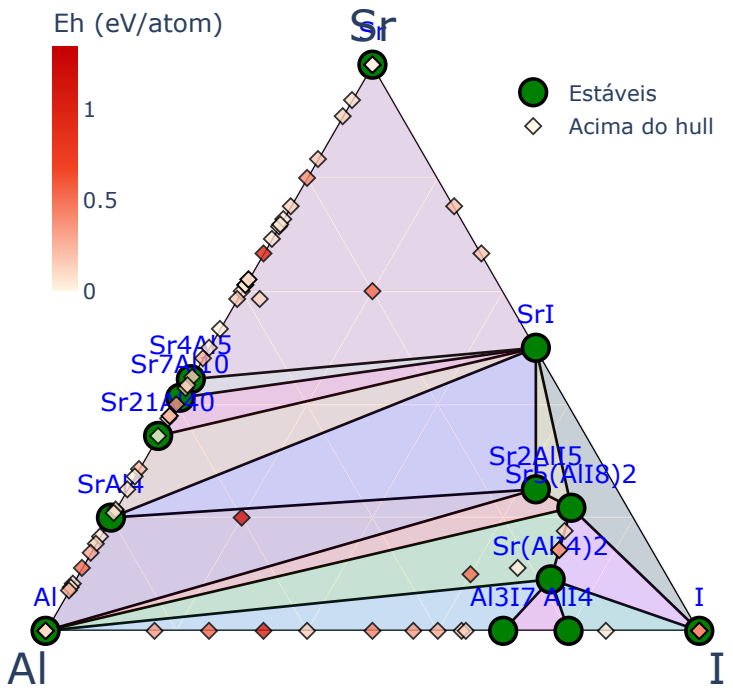
Fonte: elaborado pelo autor.

Figura A83: Diagrama de fase do sistema Sr-Al-F.



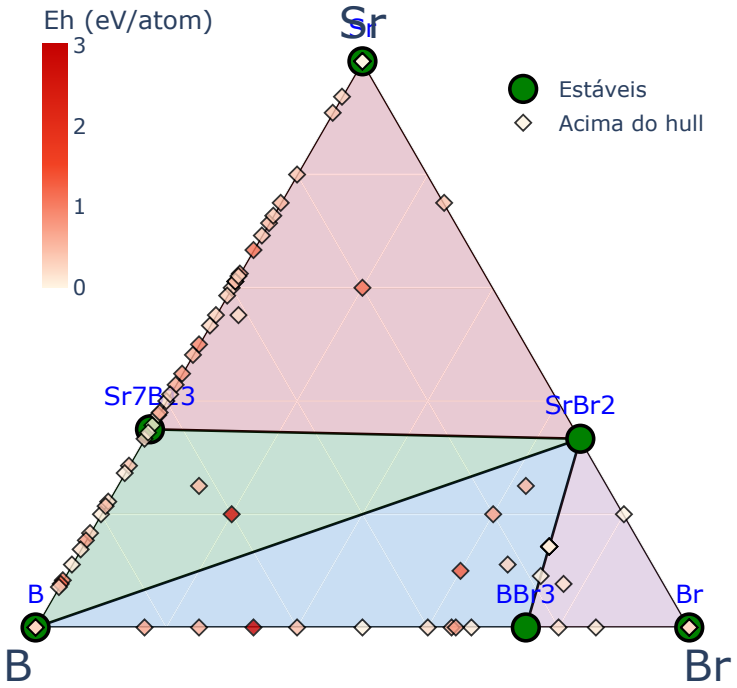
Fonte: elaborado pelo autor.

Figura A84: Diagrama de fase do sistema Sr-Al-I.



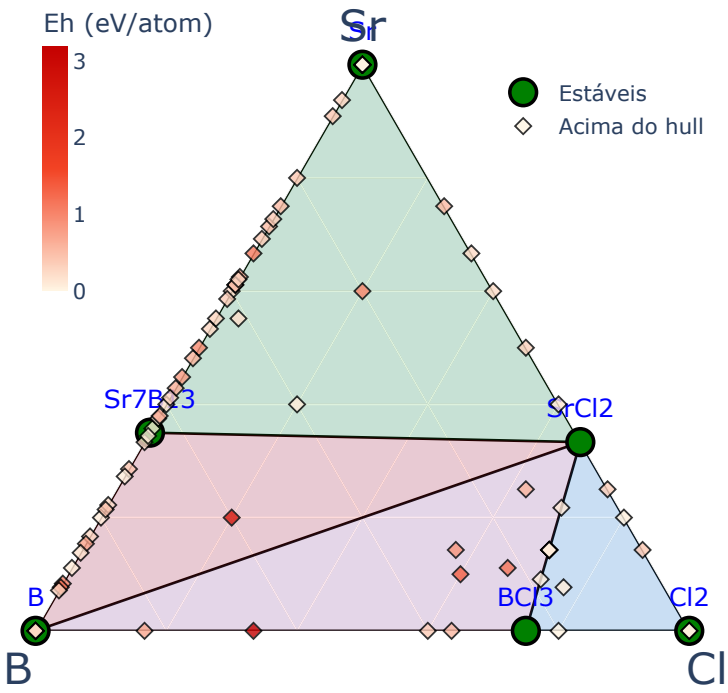
Fonte: elaborado pelo autor.

Figura A85: Diagrama de fase do sistema Sr-B-Br.



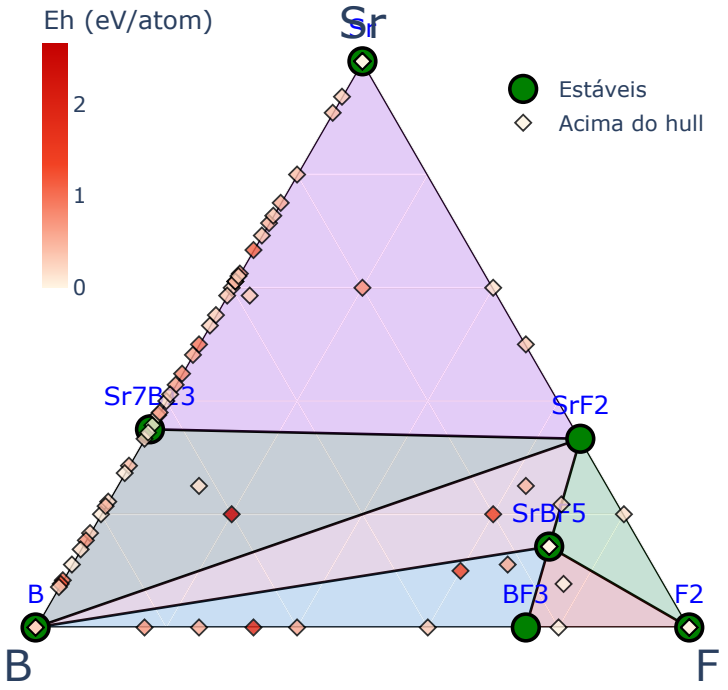
Fonte: elaborado pelo autor.

Figura A86: Diagrama de fase do sistema Sr-B-Cl.



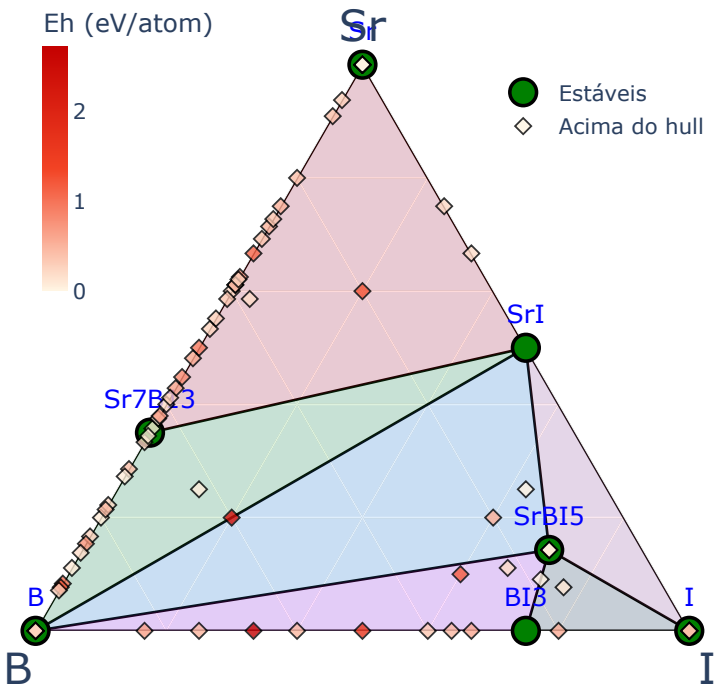
Fonte: elaborado pelo autor.

Figura A87: Diagrama de fase do sistema Sr-B-F.



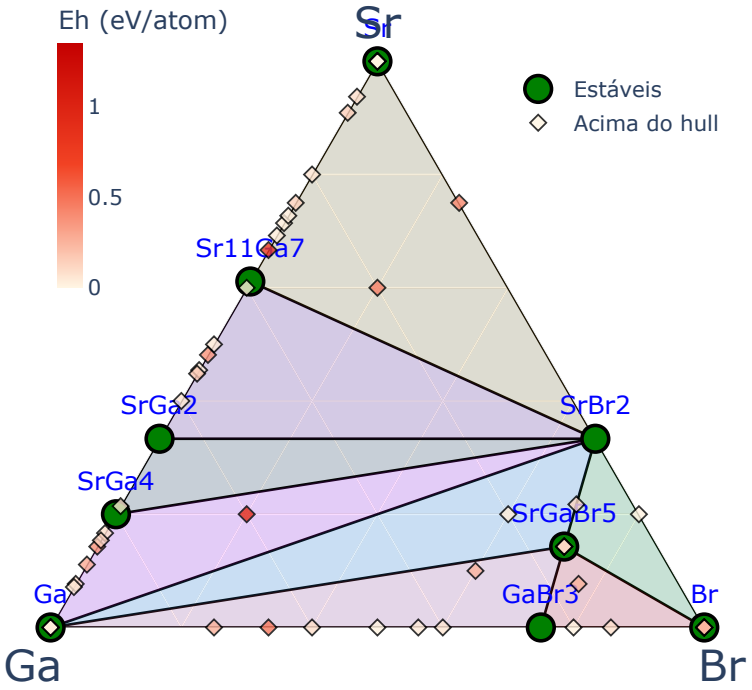
Fonte: elaborado pelo autor.

Figura A88: Diagrama de fase do sistema Sr-B-I.



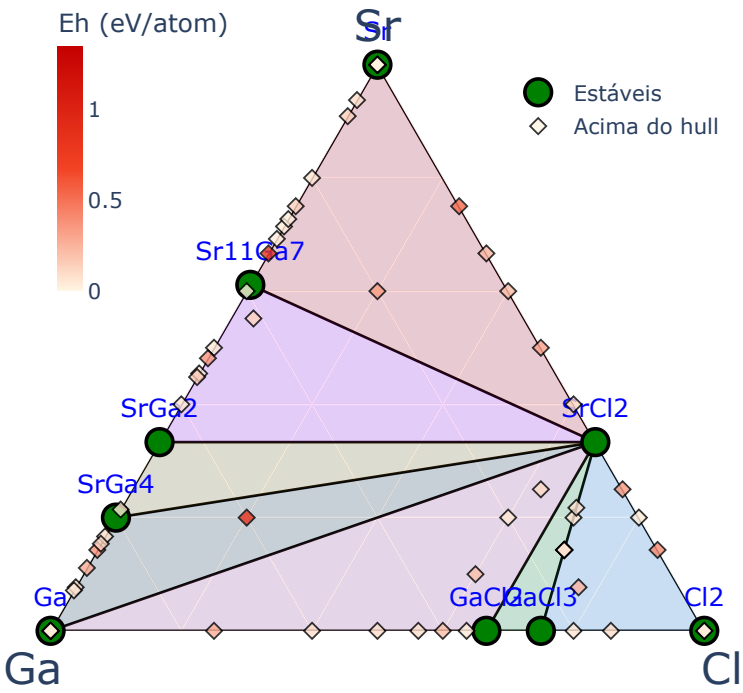
Fonte: elaborado pelo autor.

Figura A89: Diagrama de fase do sistema Sr-Ga-Br.



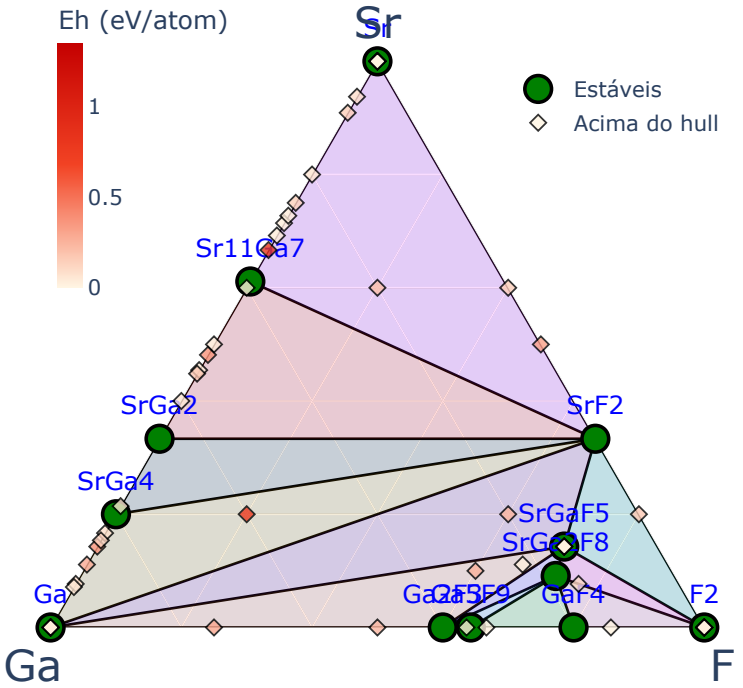
Fonte: elaborado pelo autor.

Figura A90: Diagrama de fase do sistema Sr-Ga-Cl.



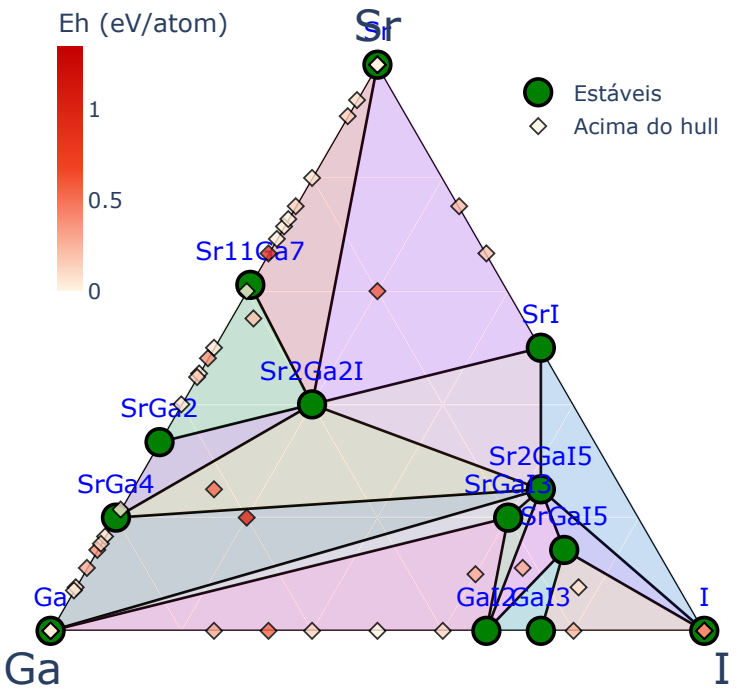
Fonte: elaborado pelo autor.

Figura A91: Diagrama de fase do sistema Sr-Ga-F.



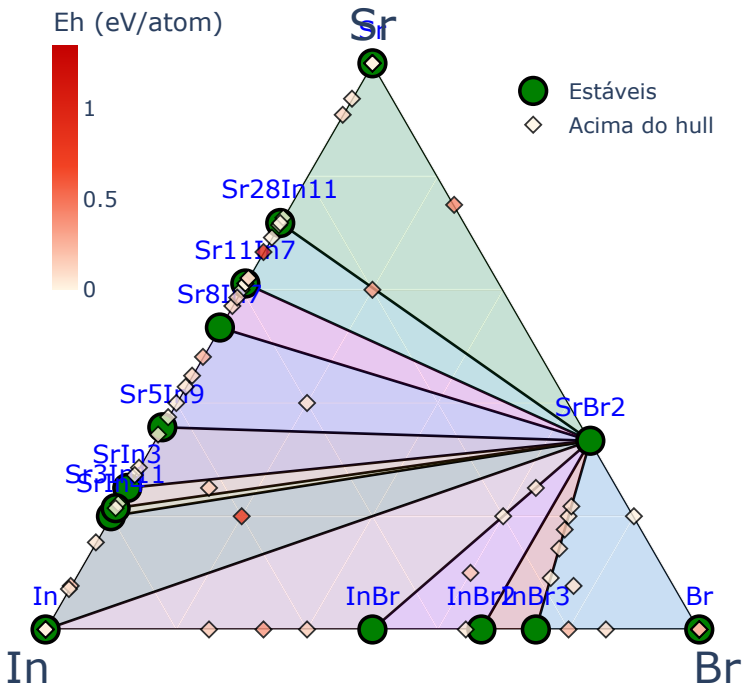
Fonte: elaborado pelo autor.

Figura A92: Diagrama de fase do sistema Sr-Ga-I.



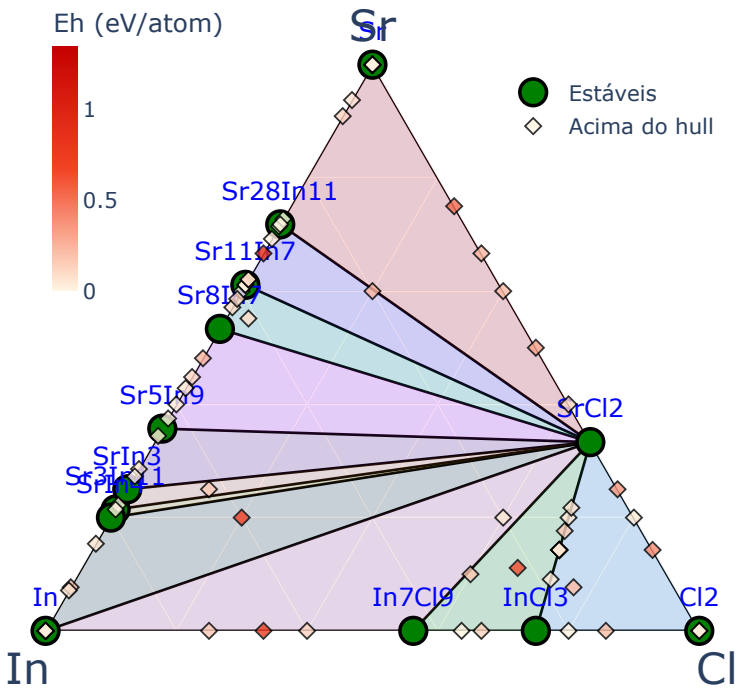
Fonte: elaborado pelo autor.

Figura A93: Diagrama de fase do sistema Sr-In-Br.



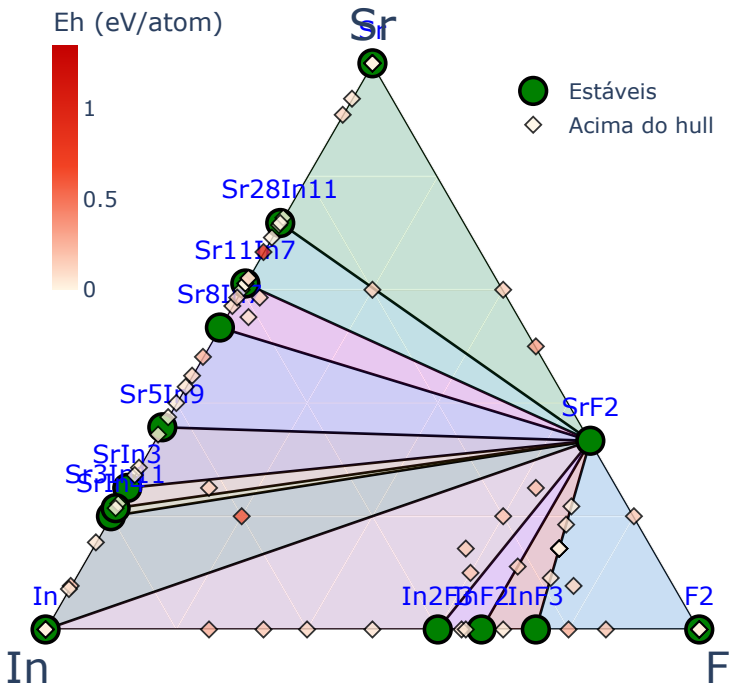
Fonte: elaborado pelo autor.

Figura A94: Diagrama de fase do sistema Sr-In-Cl.



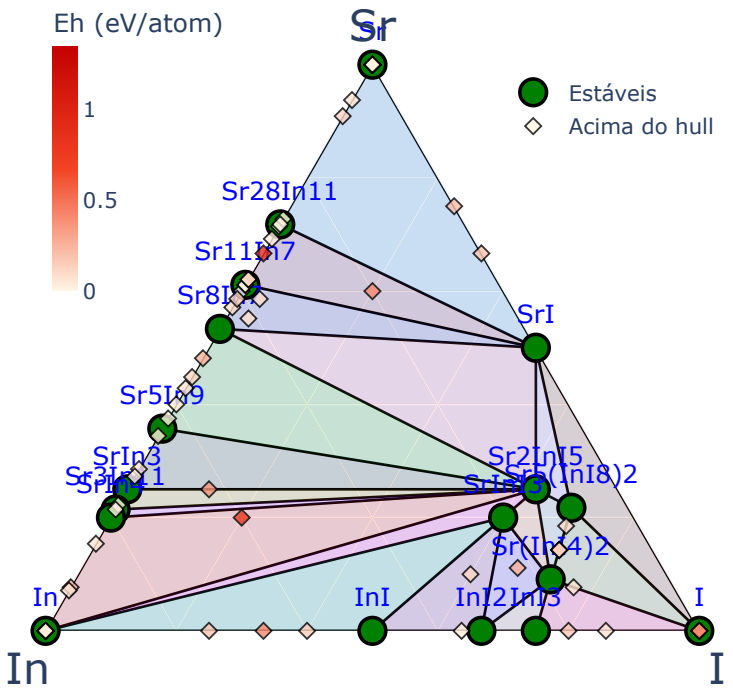
Fonte: elaborado pelo autor.

Figura A95: Diagrama de fase do sistema Sr-In-F.



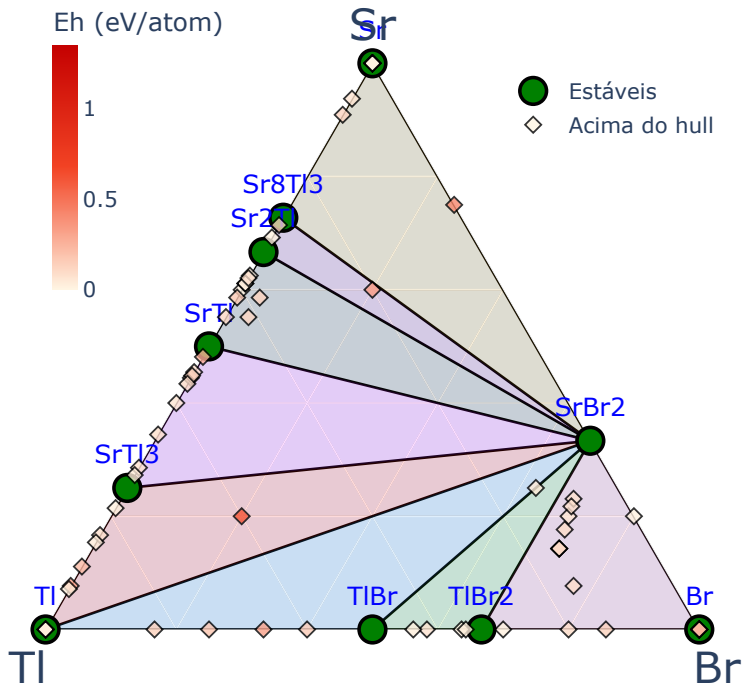
Fonte: elaborado pelo autor.

Figura A96: Diagrama de fase do sistema Sr-In-I.



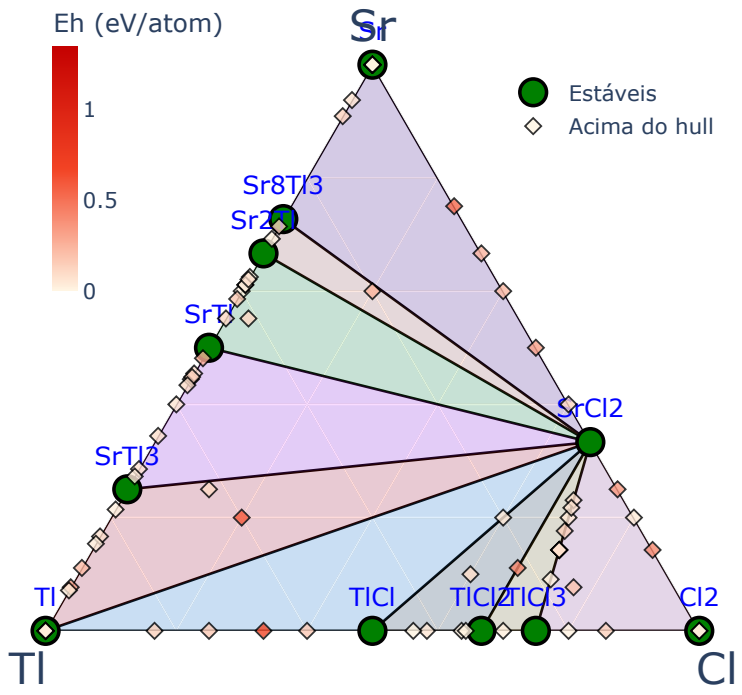
Fonte: elaborado pelo autor.

Figura A97: Diagrama de fase do sistema Sr-Tl-Br.



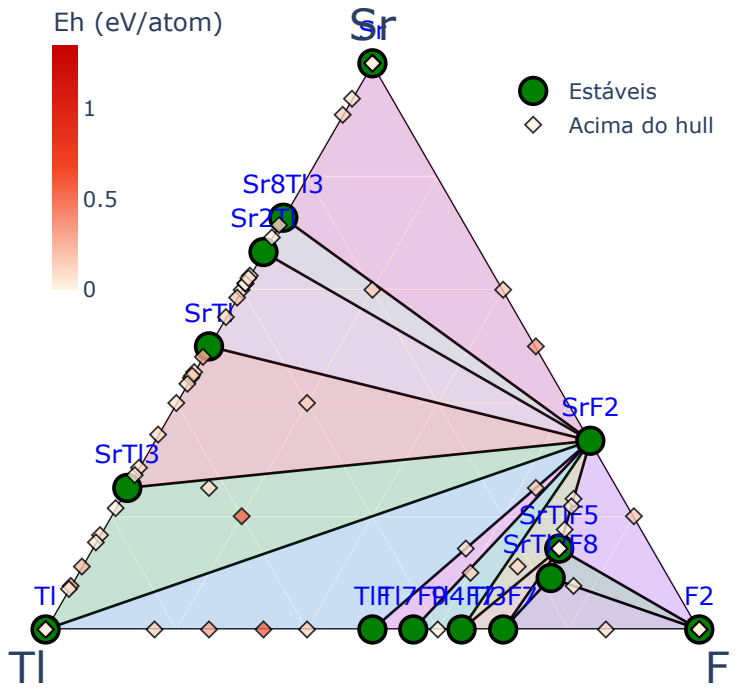
Fonte: elaborado pelo autor.

Figura A98: Diagrama de fase do sistema Sr-Tl-Cl.



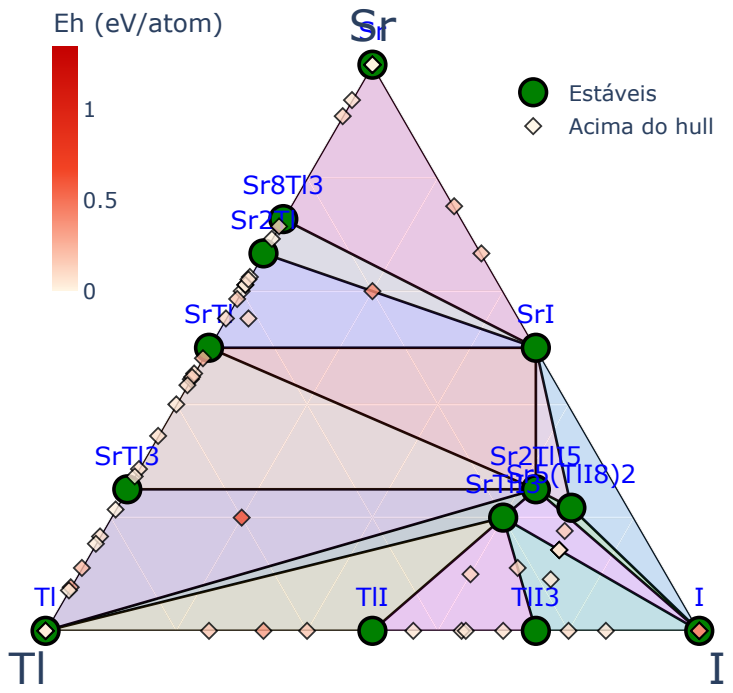
Fonte: elaborado pelo autor.

Figura A99: Diagrama de fase do sistema Sr-Tl-F.



Fonte: elaborado pelo autor.

Figura A100: Diagrama de fase do sistema Sr-Tl-I.

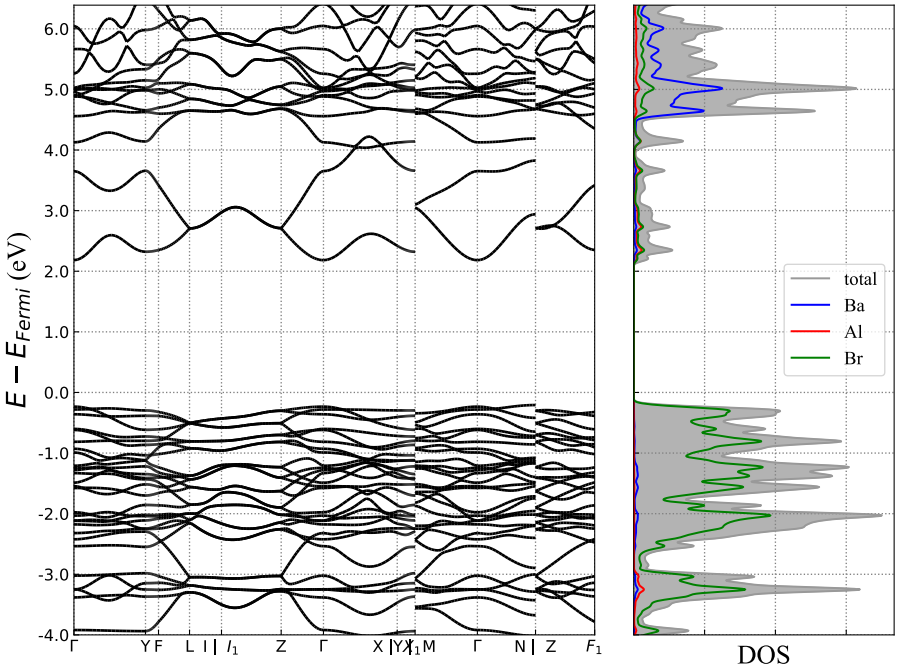


Fonte: elaborado pelo autor.

Apêndice B

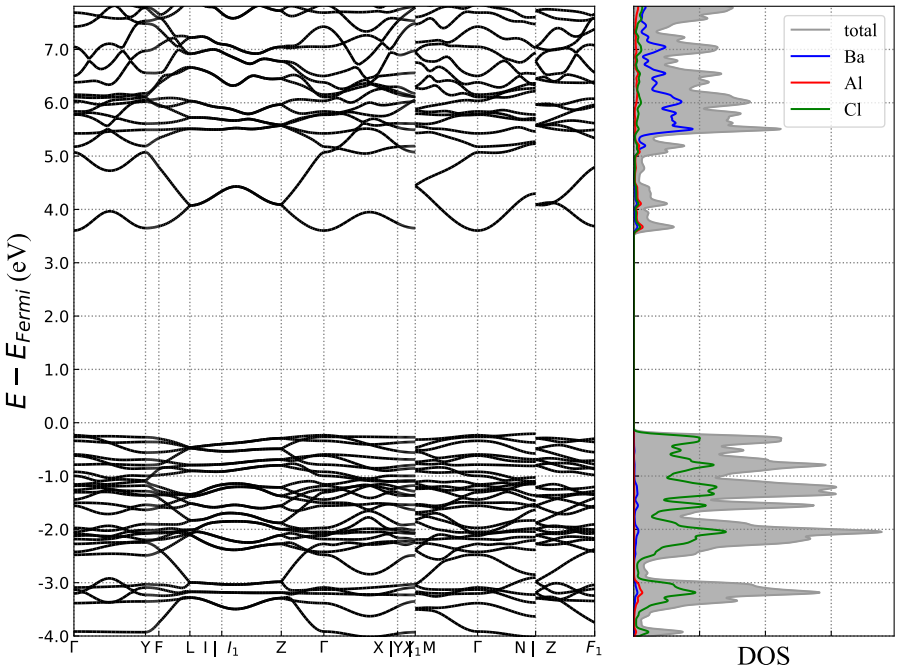
Neste apêndice são apresentadas as estruturas de bandas e a densidade de estados eletrônicos (DOS) para os materiais alvo $\alpha-CaAlF_5$ (grupo espacial de número 15), referentes ao estudo de caso II, discutido no Capítulo 5. Estes resultados foram obtidos por simulações DFT PBE-GGA.

Figura B1: Estrutura de bandas e DOS do BaAlBr5.
Bandgap indireto de $E_g = 2.39$ eV.



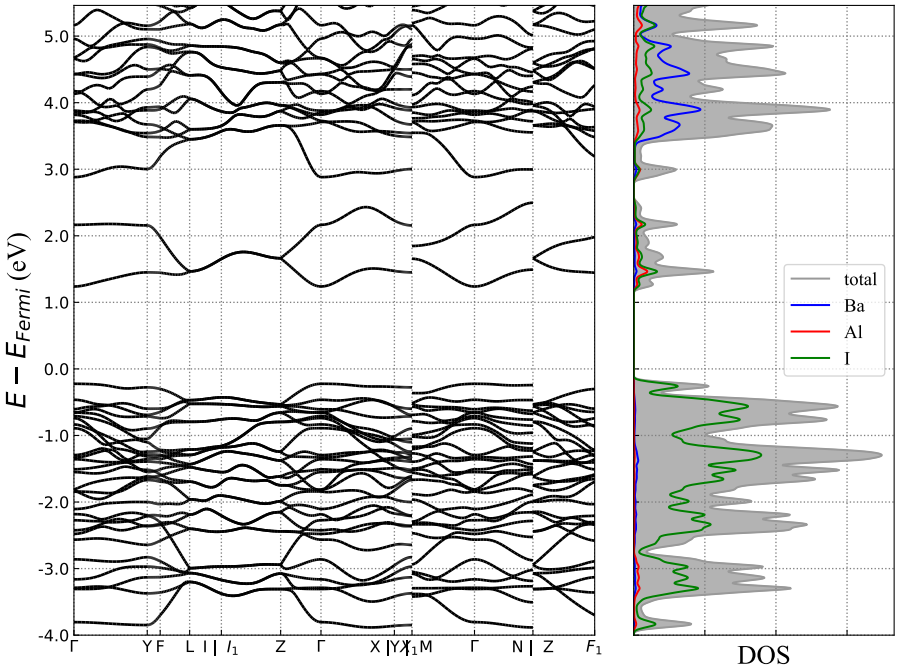
Fonte: elaborado pelo autor.

Figura B2: Estrutura de bandas e DOS do BaAlCl5.
Bandgap indireto de $E_g = 3.81$ eV.



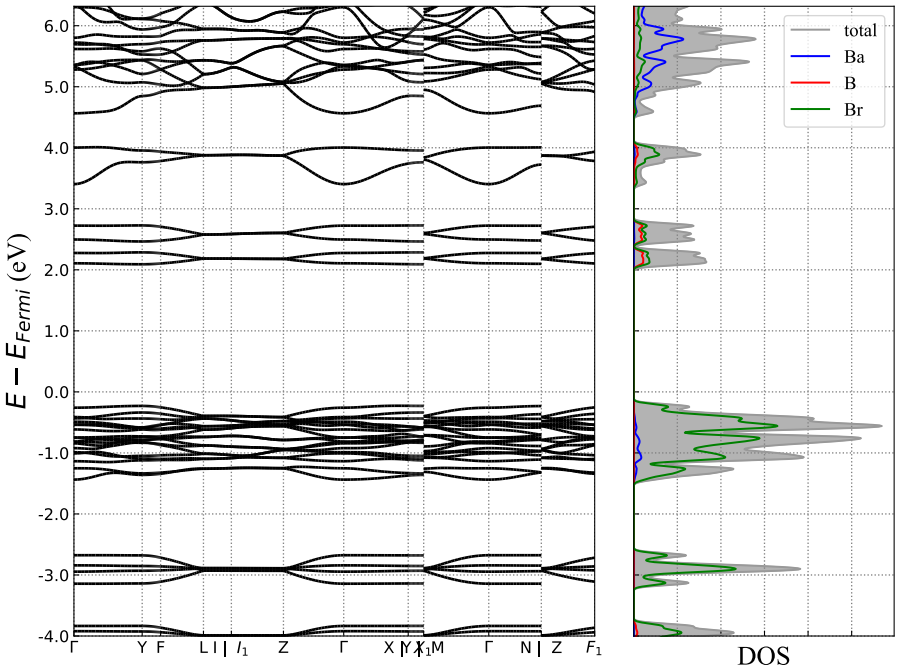
Fonte: elaborado pelo autor.

Figura B3: Estrutura de bandas e DOS do BaAlI5.
Bandgap direto de $E_g = 1.46$ eV.



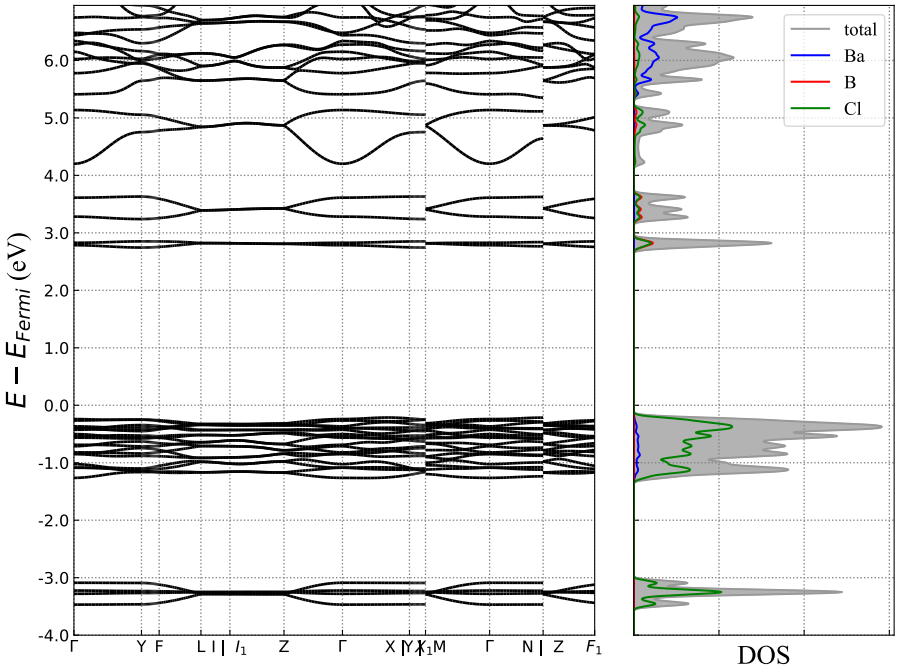
Fonte: elaborado pelo autor.

Figura B4: Estrutura de bandas e DOS do BaBBr5.
Bandgap indireto de $E_g = 2.32$ eV.



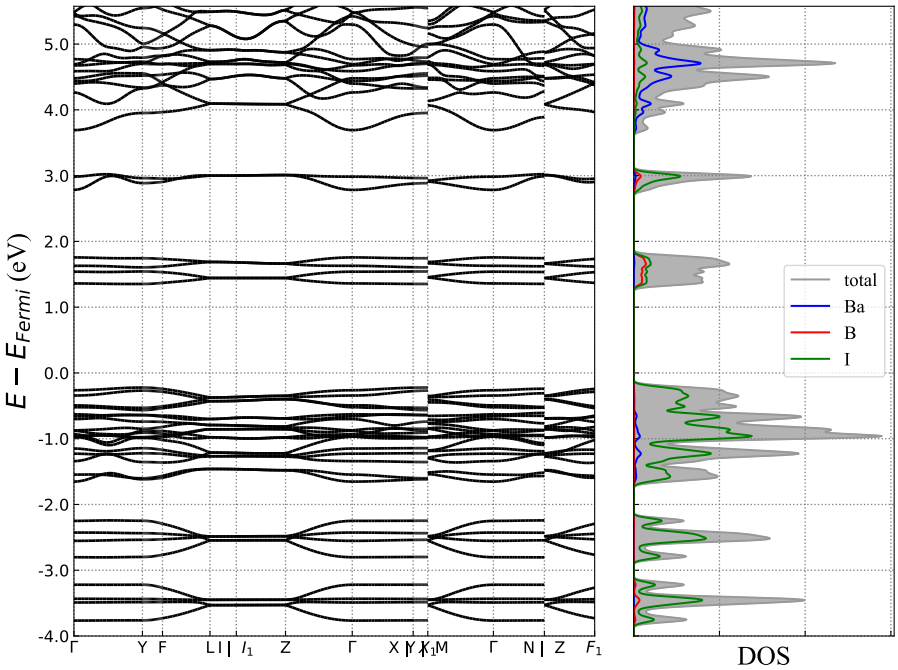
Fonte: elaborado pelo autor.

Figura B5: Estrutura de bandas e DOS do BaBCl5.
Bandgap indireto de $E_g = 2.96$ eV.



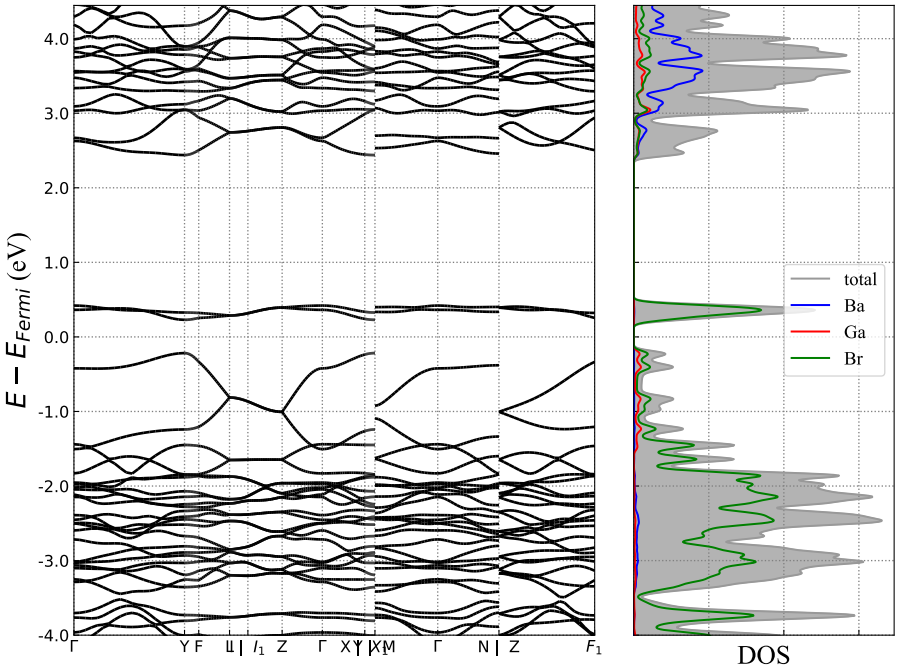
Fonte: elaborado pelo autor.

Figura B6: Estrutura de bandas e DOS do BaBI5.
Bandgap indireto de $E_g = 1.57$ eV.



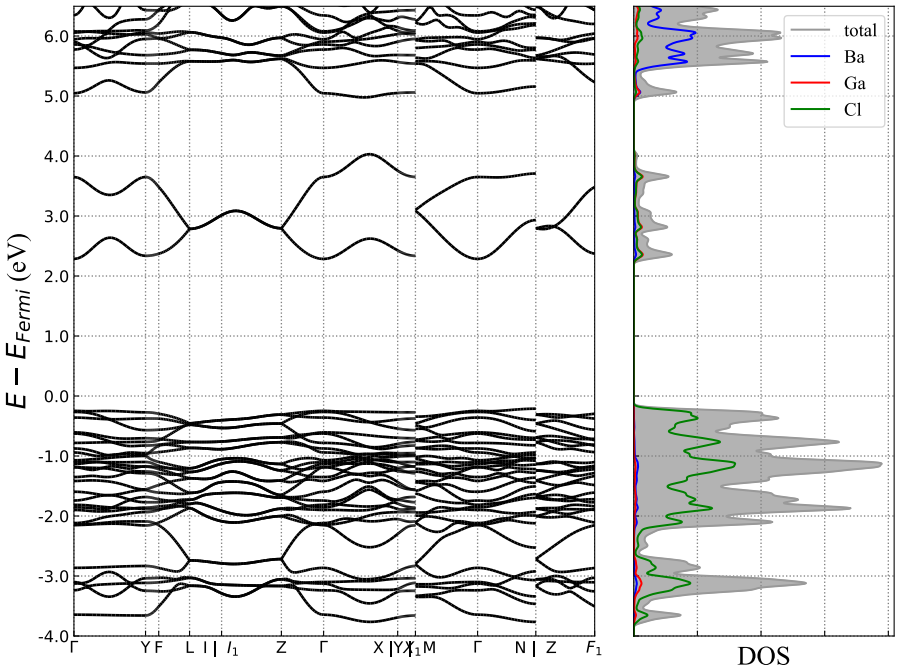
Fonte: elaborado pelo autor.

Figura B7: Estrutura de bandas e DOS do BaGaBr5.
Bandgap direto de $E_g = 0.45$ eV.



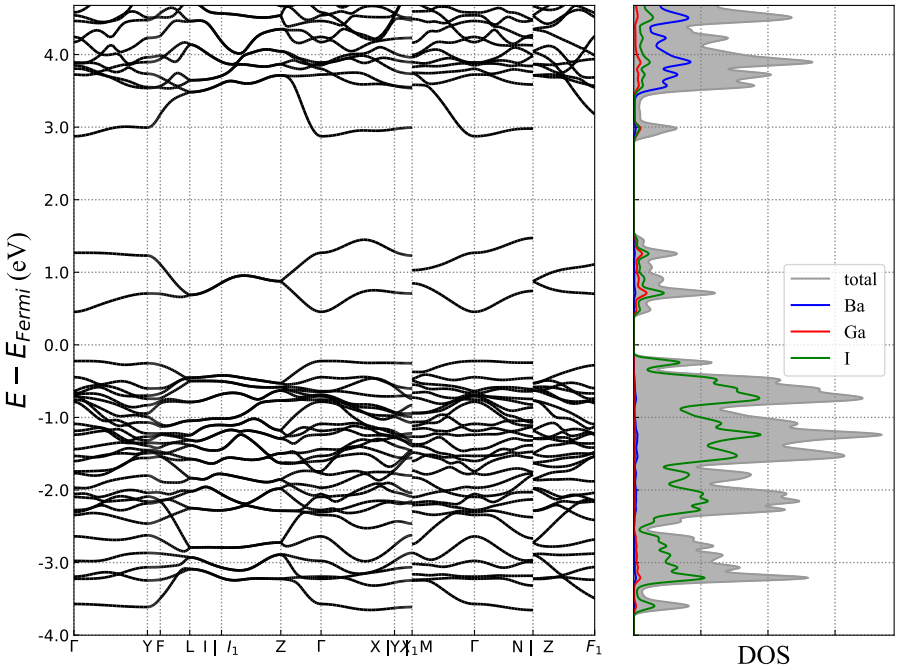
Fonte: elaborado pelo autor.

Figura B8: Estrutura de bandas e DOS do BaGaCl5.
Bandgap indireto de $E_g = 2.5$ eV.



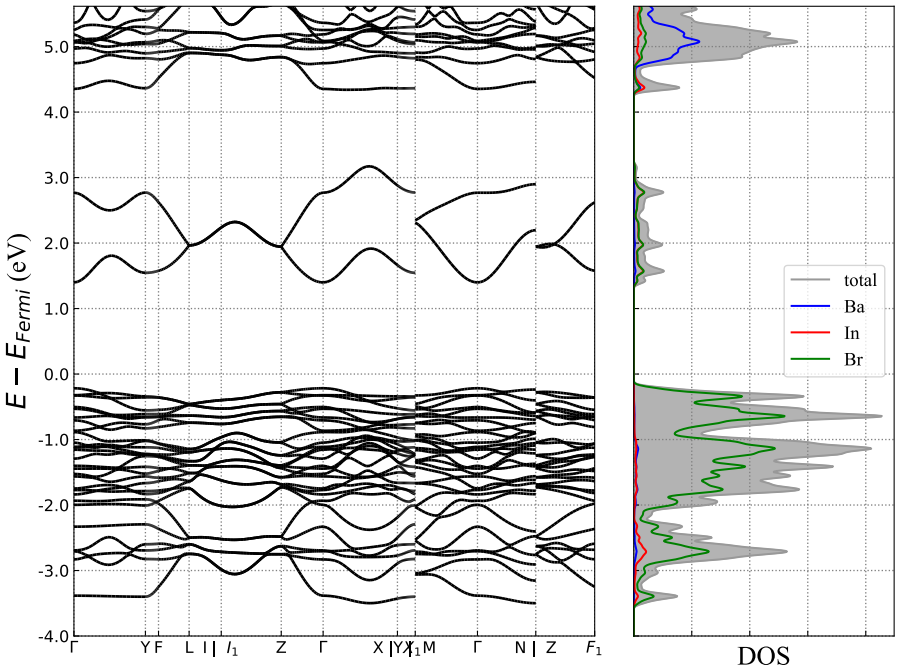
Fonte: elaborado pelo autor.

Figura B9: Estrutura de bandas e DOS do BaGaI5.
Bandgap indireto de $E_g = 0.68$ eV.



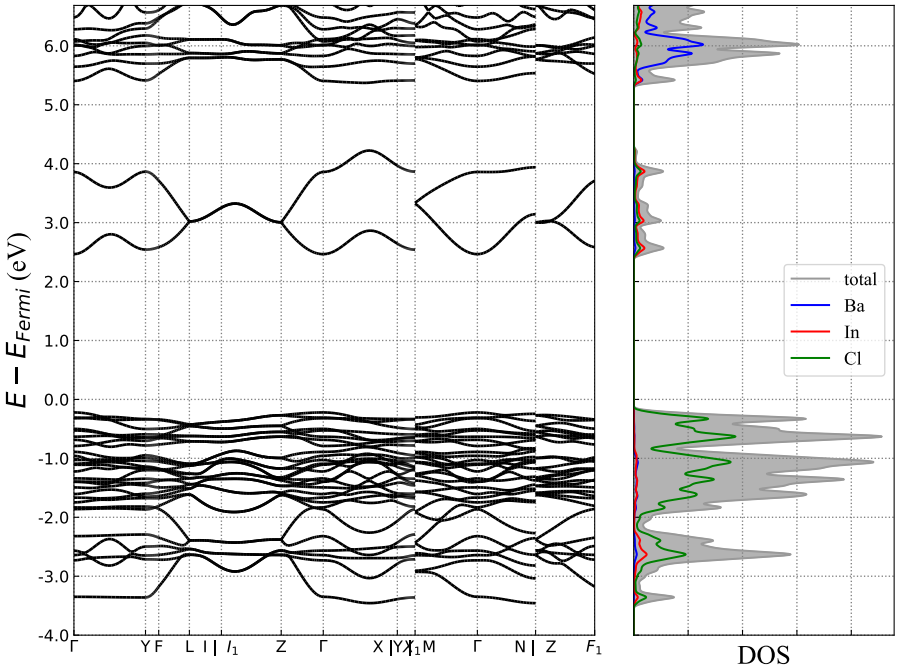
Fonte: elaborado pelo autor.

Figura B10: Estrutura de bandas e DOS do BaInBr5.
Bandgap direto de $E_g = 1.62$ eV.



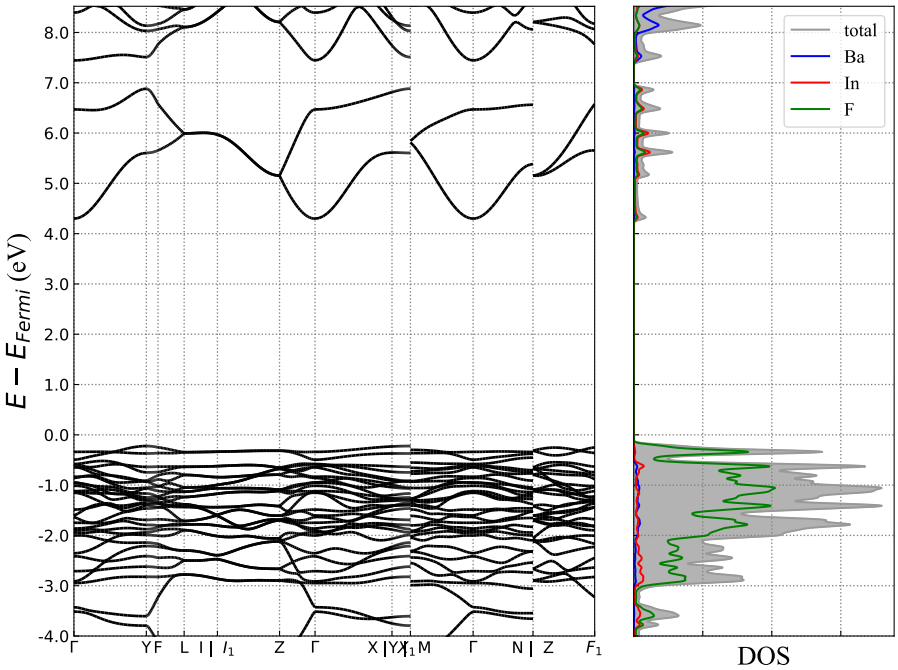
Fonte: elaborado pelo autor.

Figura B11: Estrutura de bandas e DOS do BaInCl5.
Bandgap direto de $E_g = 2.69$ eV.



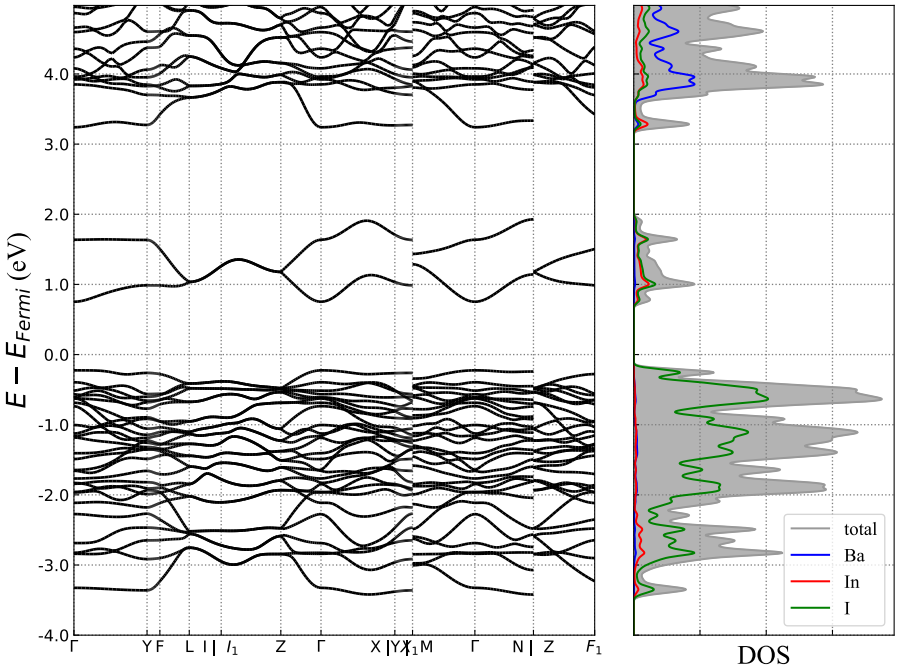
Fonte: elaborado pelo autor.

Figura B12: Estrutura de bandas e DOS do BaInF5.
Bandgap indireto de $E_g = 4.52$ eV.



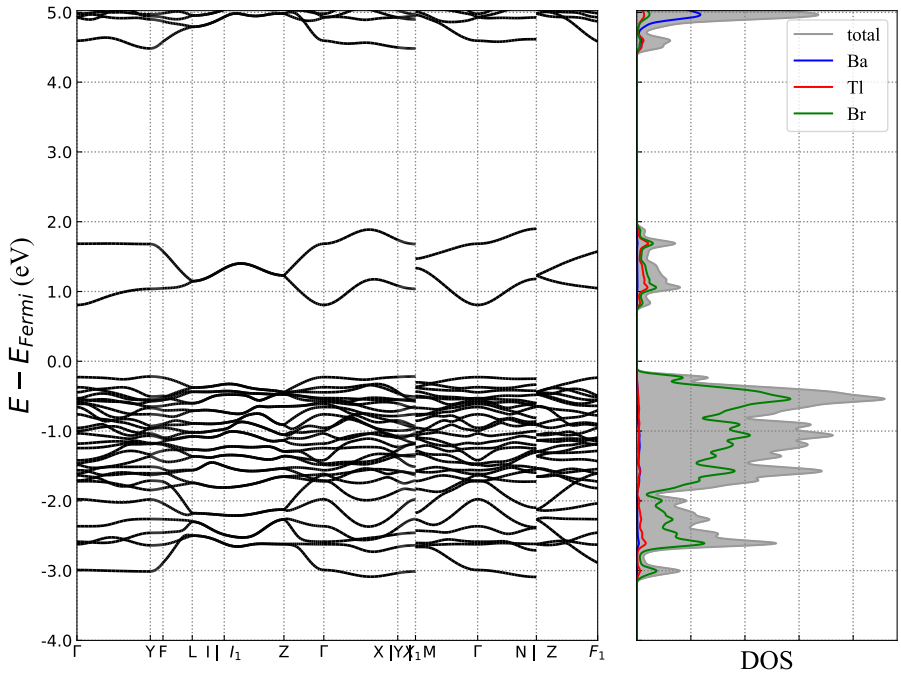
Fonte: elaborado pelo autor.

Figura B13: Estrutura de bandas e DOS do BaInI5.
Bandgap direto de $E_g = 0.98$ eV.



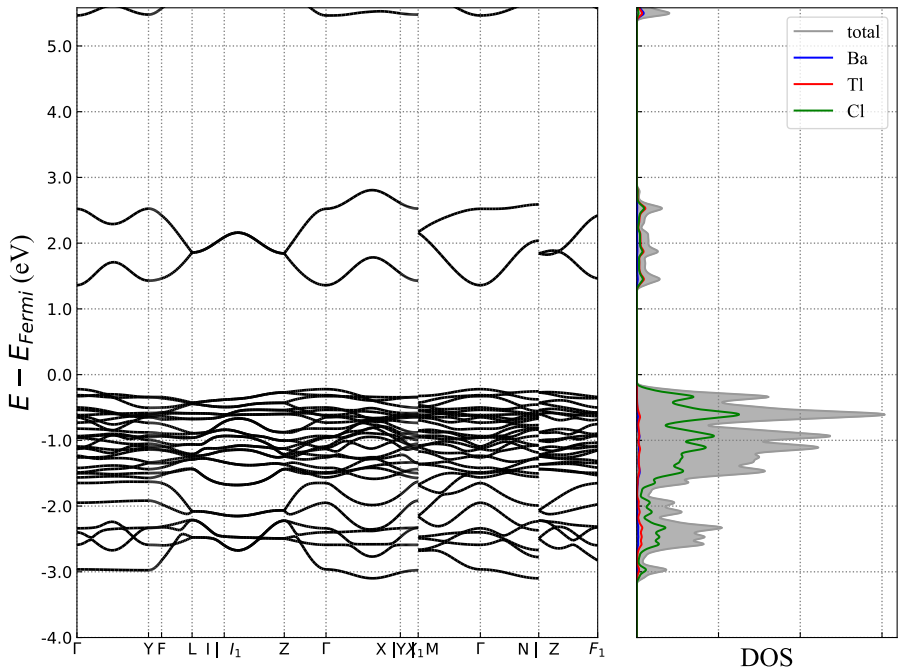
Fonte: elaborado pelo autor.

Figura B14: Estrutura de bandas e DOS do BaTlBr5.
Bandgap indireto de $E_g = 1.02$ eV.



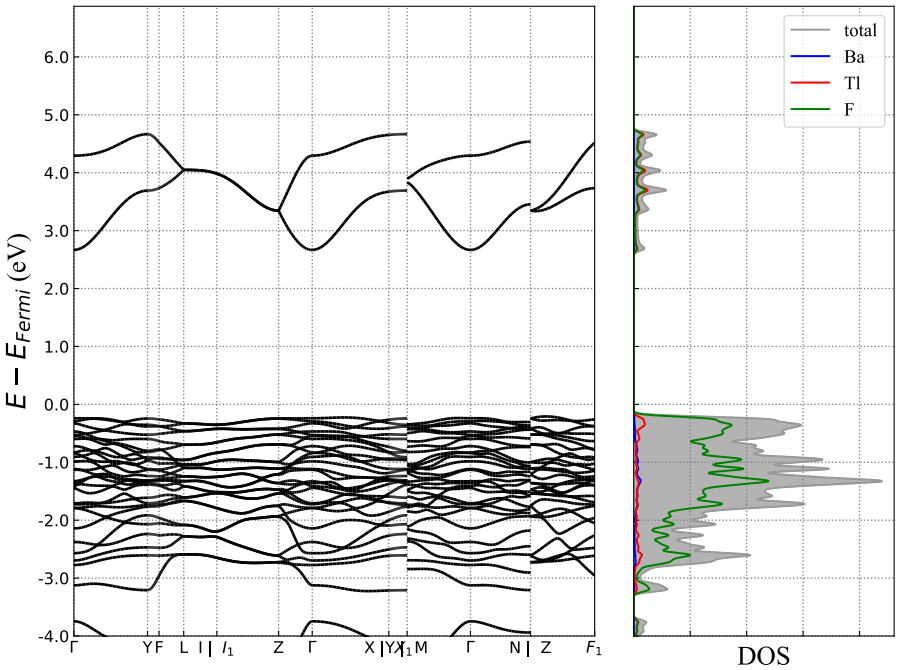
Fonte: elaborado pelo autor.

Figura B15: Estrutura de bandas e DOS do BaTlCl5.
Bandgap direto de $E_g = 1.58$ eV.



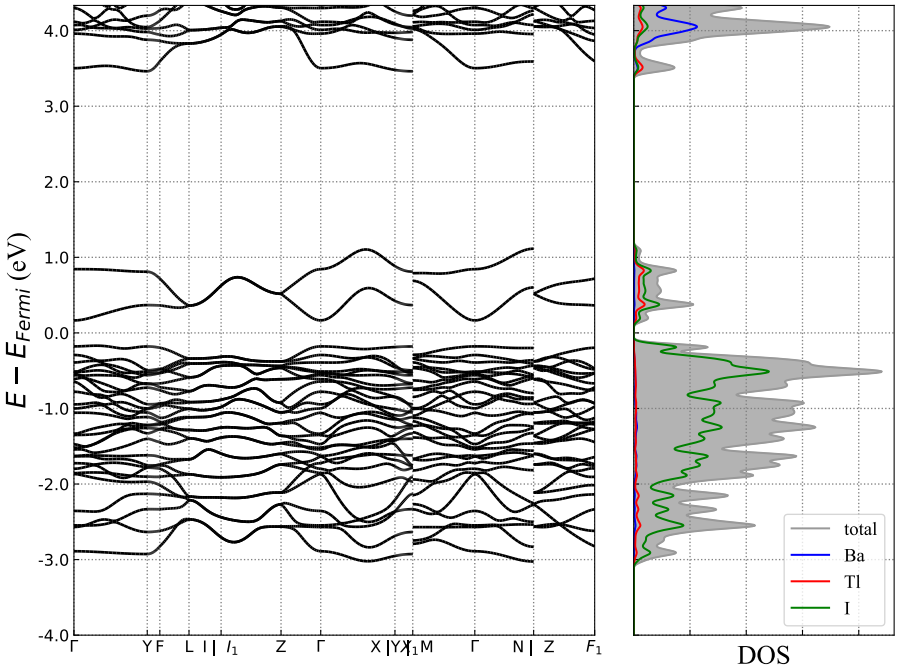
Fonte: elaborado pelo autor.

Figura B16: Estrutura de bandas e DOS do BaTlF5.
Bandgap indireto de $E_g = 2.87$ eV.



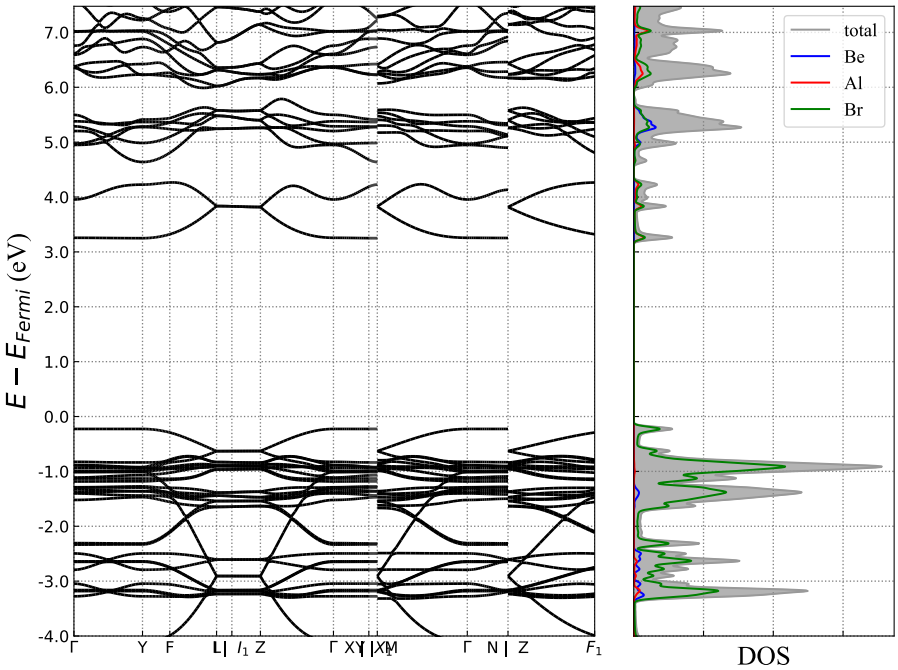
Fonte: elaborado pelo autor.

Figura B17: Estrutura de bandas e DOS do BaTlI5.
Bandgap indireto de $E_g = 0.33$ eV.



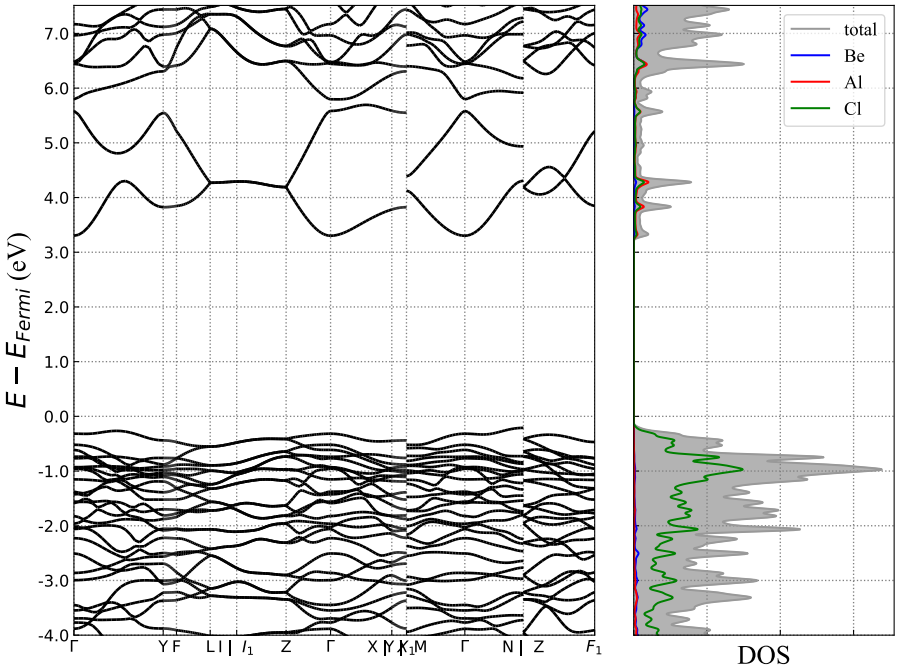
Fonte: elaborado pelo autor.

Figura B18: Estrutura de bandas e DOS do BeAlBr5.
Bandgap indireto de $E_g = 3.48$ eV.



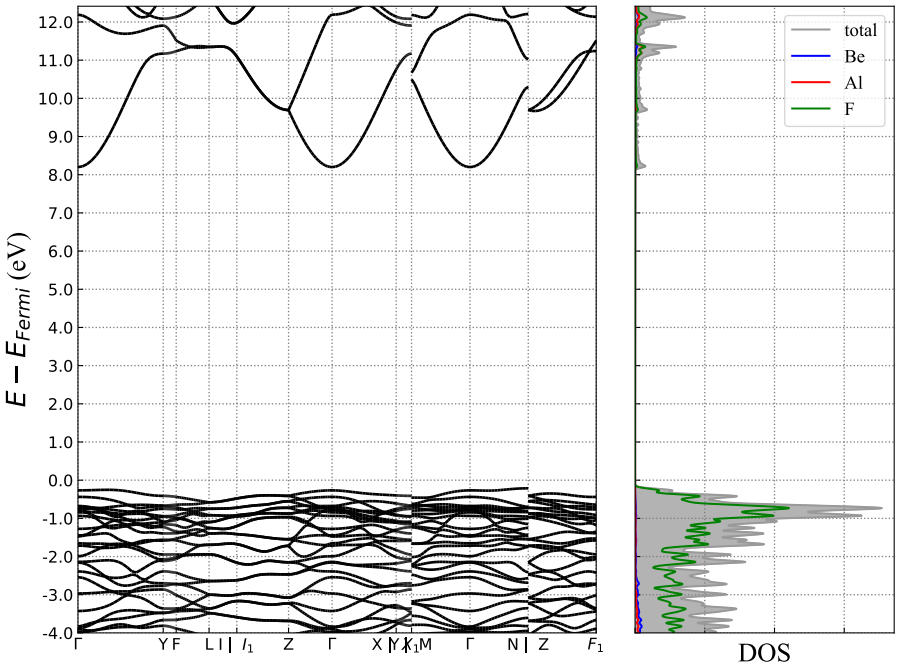
Fonte: elaborado pelo autor.

Figura B19: Estrutura de bandas e DOS do BeAlCl5.
Bandgap indireto de $E_g = 3.51$ eV.



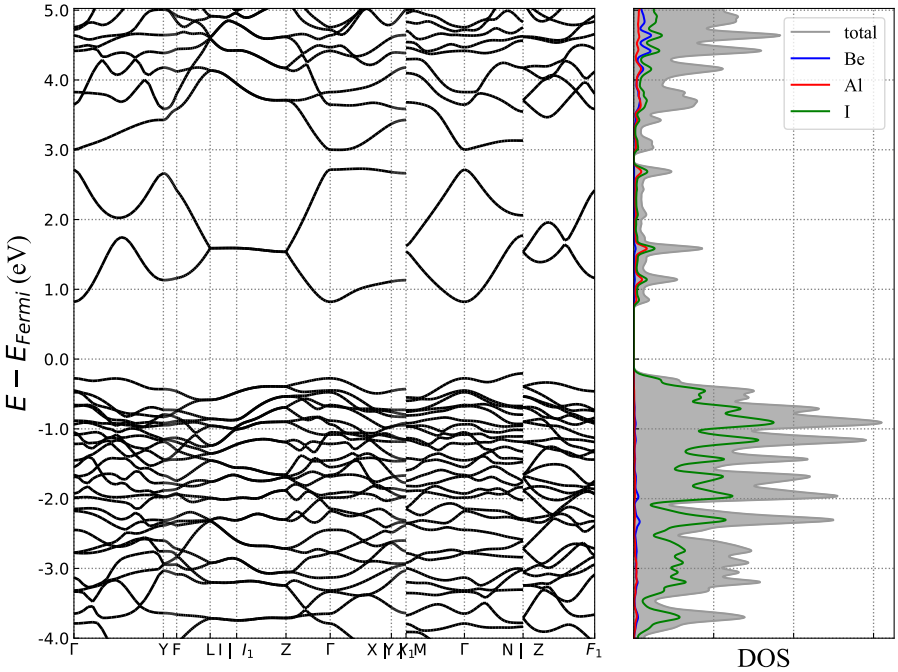
Fonte: elaborado pelo autor.

Figura B20: Estrutura de bandas e DOS do BeAlF5.
Bandgap indireto de $E_g = 8.41$ eV.



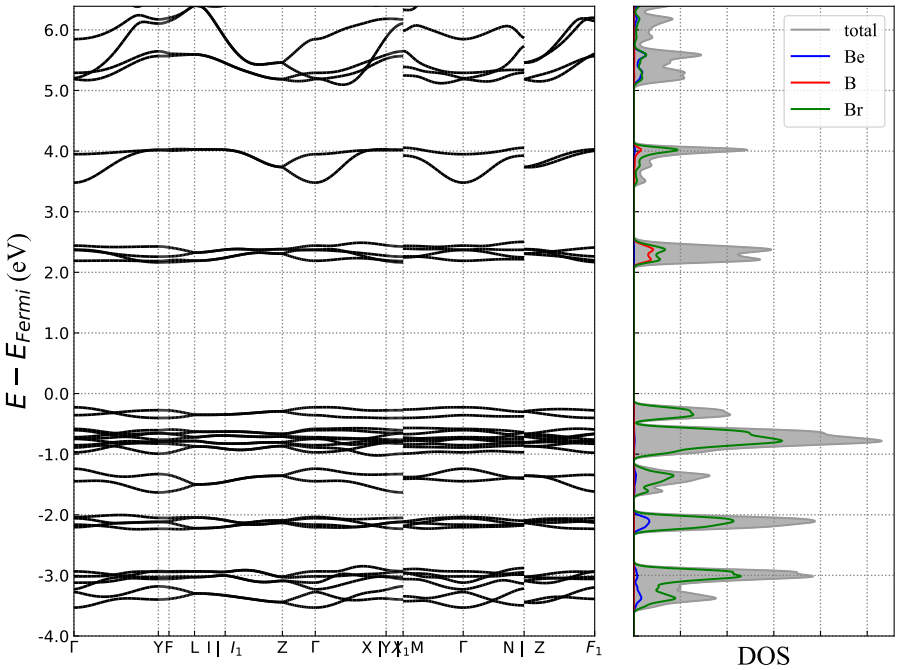
Fonte: elaborado pelo autor.

Figura B21: Estrutura de bandas e DOS do BeAlI5.
Bandgap indireto de $E_g = 1.03$ eV.



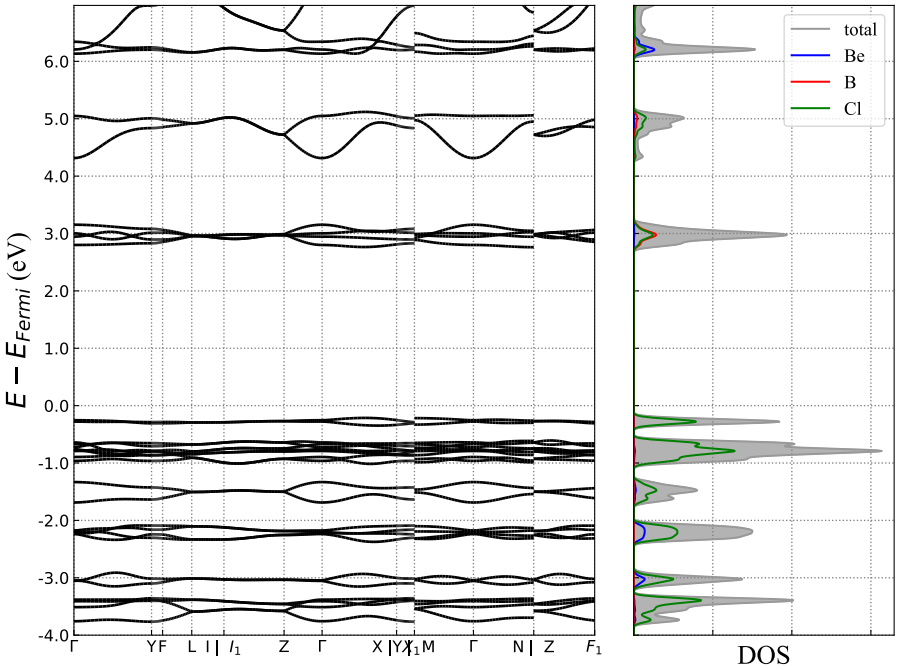
Fonte: elaborado pelo autor.

Figura B22: Estrutura de bandas e DOS do BeBBr5.
Bandgap indireto de $E_g = 2.39$ eV.



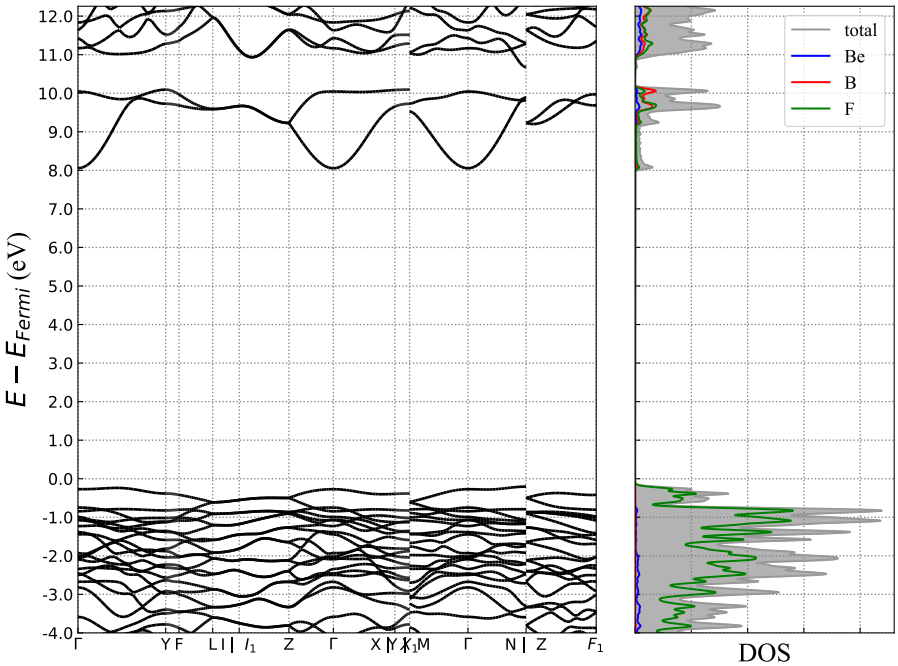
Fonte: elaborado pelo autor.

Figura B23: Estrutura de bandas e DOS do BeBCl5.
Bandgap indireto de $E_g = 2.98$ eV.



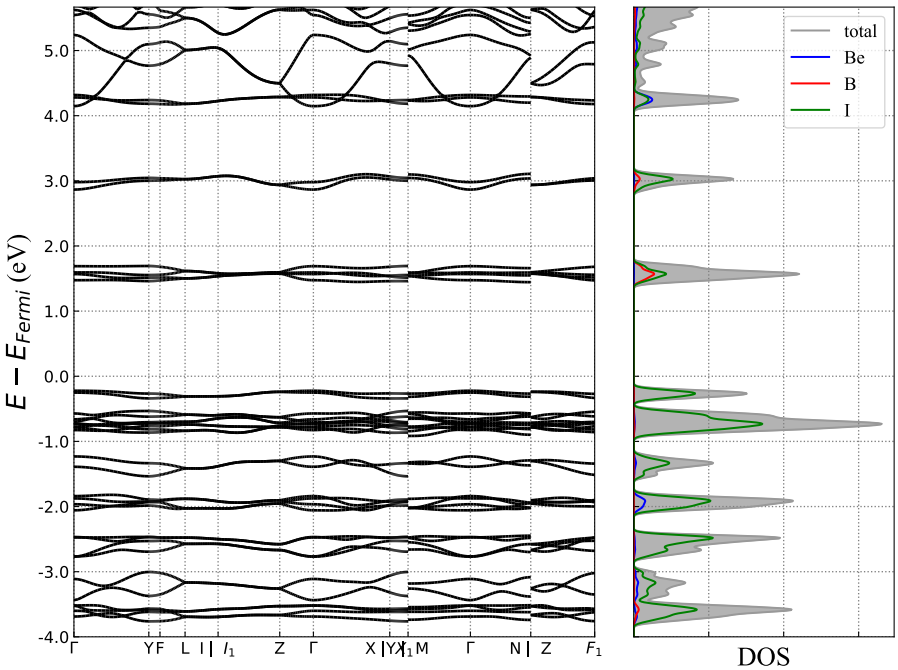
Fonte: elaborado pelo autor.

Figura B24: Estrutura de bandas e DOS do BeBF5.
Bandgap indireto de $E_g = 8.26$ eV.



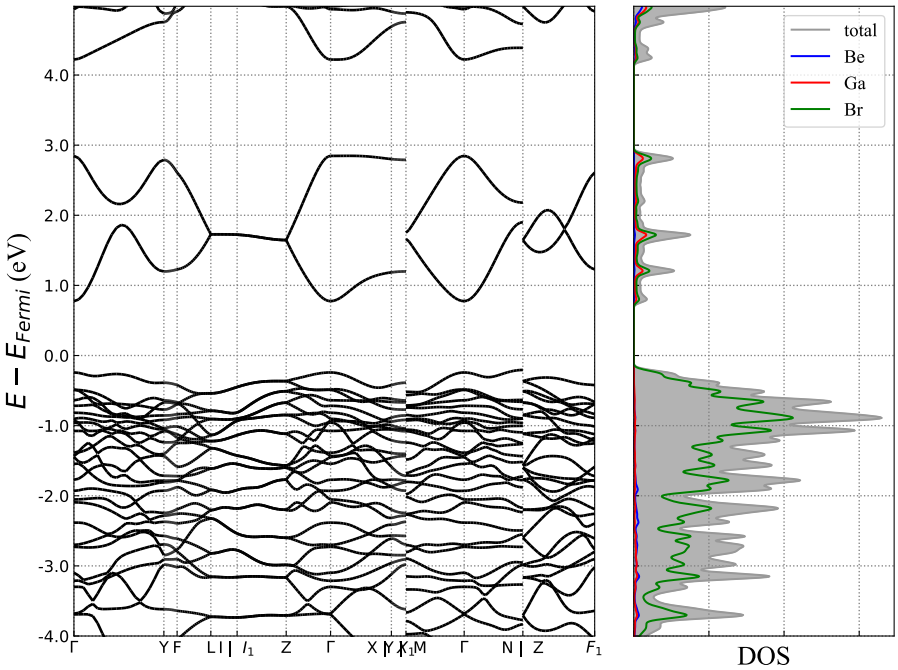
Fonte: elaborado pelo autor.

Figura B25: Estrutura de bandas e DOS do BeBI5.
Bandgap indireto de $E_g = 1.67$ eV.



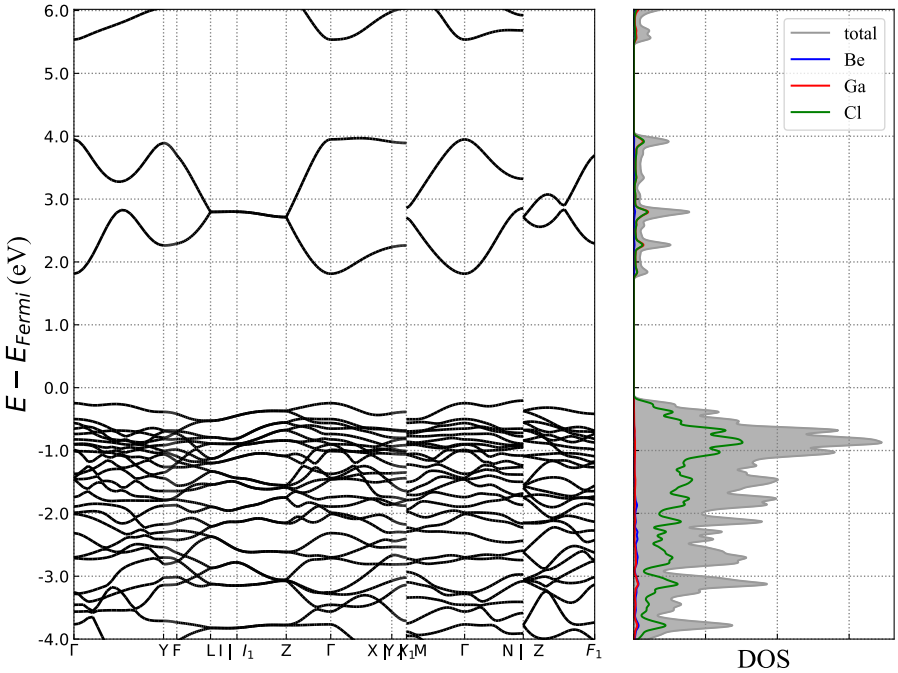
Fonte: elaborado pelo autor.

Figura B26: Estrutura de bandas e DOS do BeGaBr5.
Bandgap indireto de $E_g = 0.98$ eV.



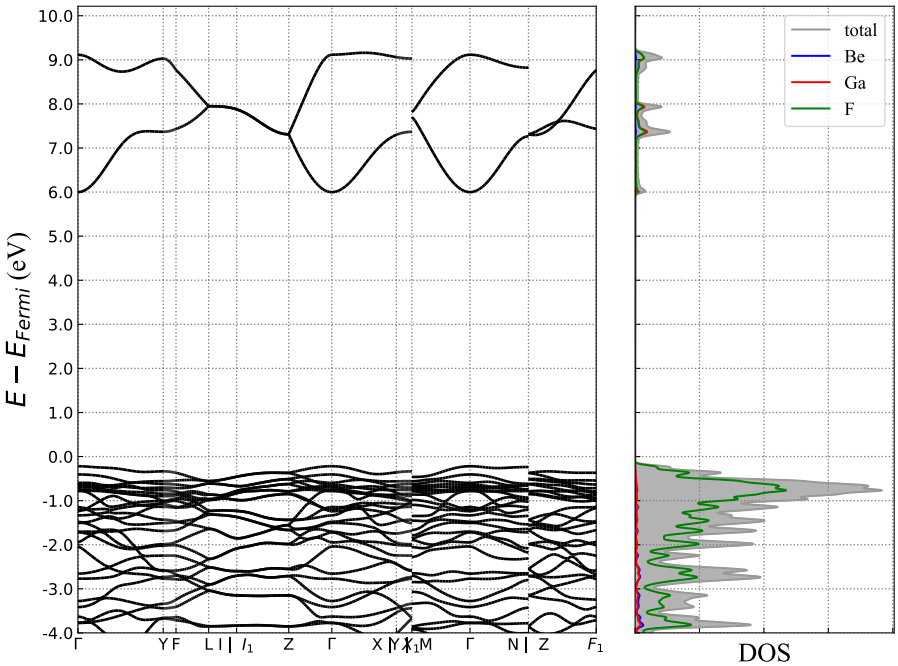
Fonte: elaborado pelo autor.

Figura B27: Estrutura de bandas e DOS do BeGaCl5.
Bandgap indireto de $E_g = 2.02$ eV.



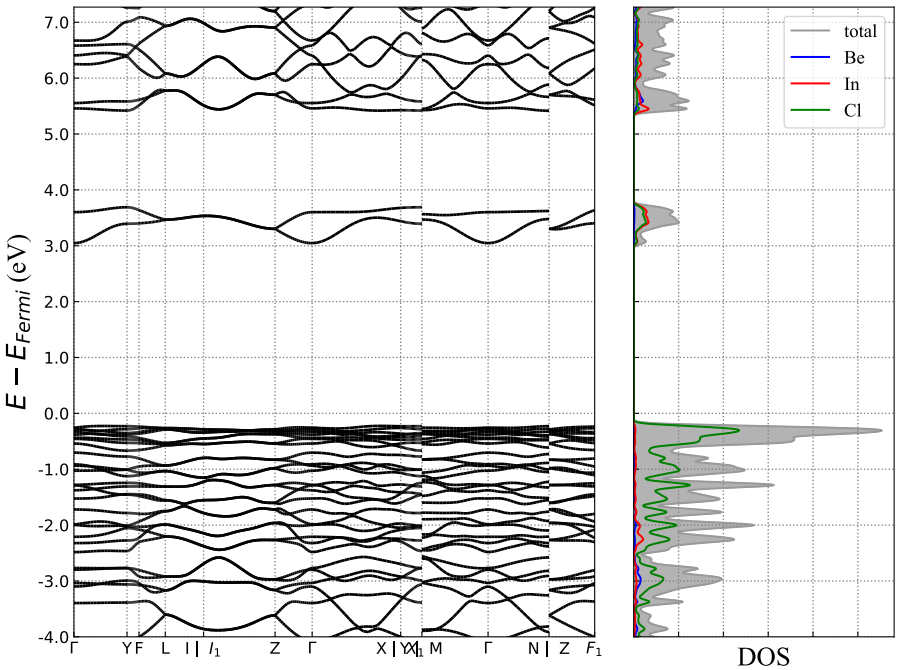
Fonte: elaborado pelo autor.

Figura B28: Estrutura de bandas e DOS do BeGaF5.
Bandgap direto de $E_g = 6.22$ eV.



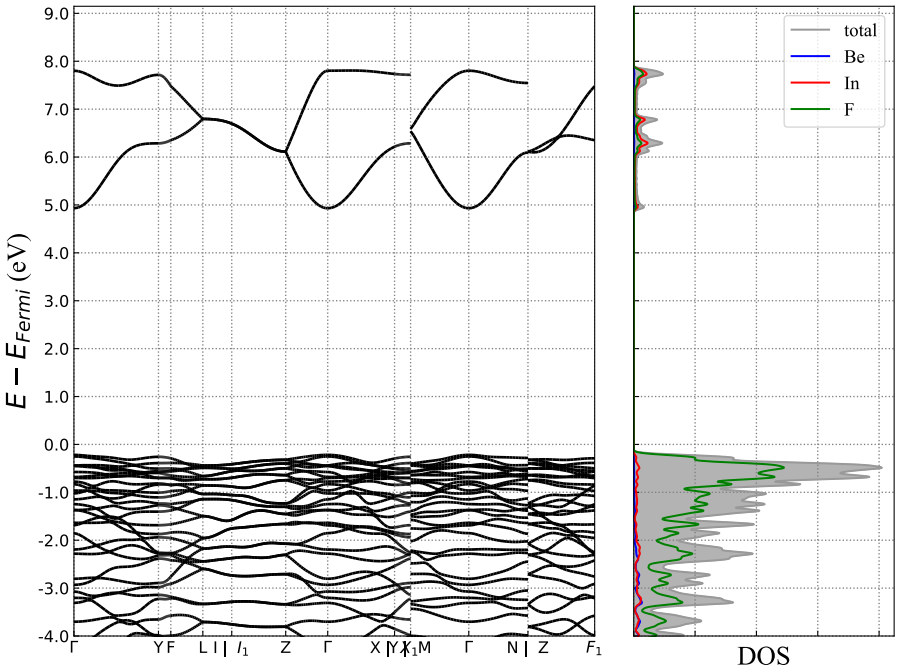
Fonte: elaborado pelo autor.

Figura B29: Estrutura de bandas e DOS do BeInCl5.
Bandgap indireto de $E_g = 3.27$ eV.



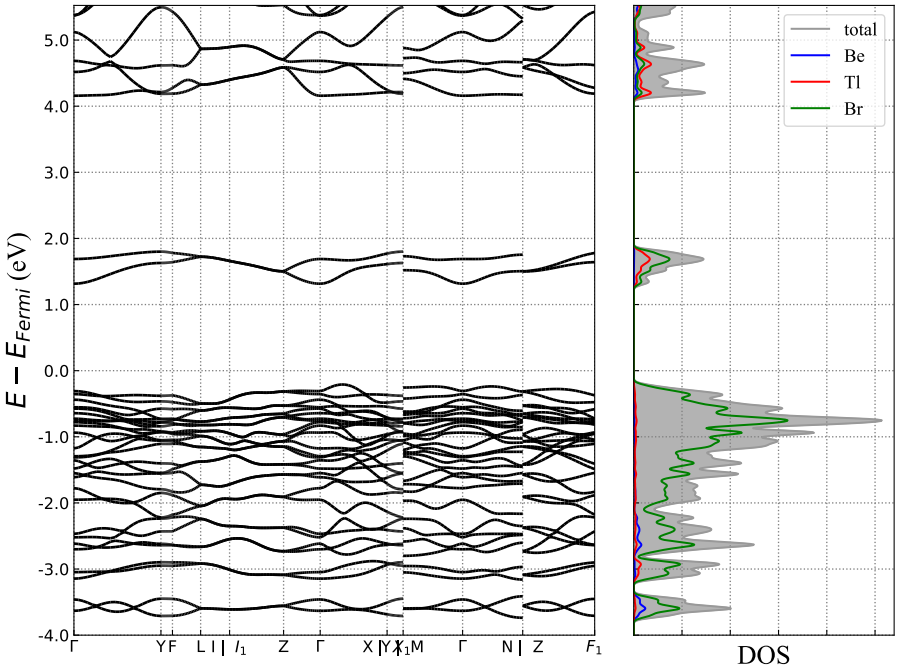
Fonte: elaborado pelo autor.

Figura B30: Estrutura de bandas e DOS do BeInF5.
Bandgap direto de $E_g = 5.15$ eV.



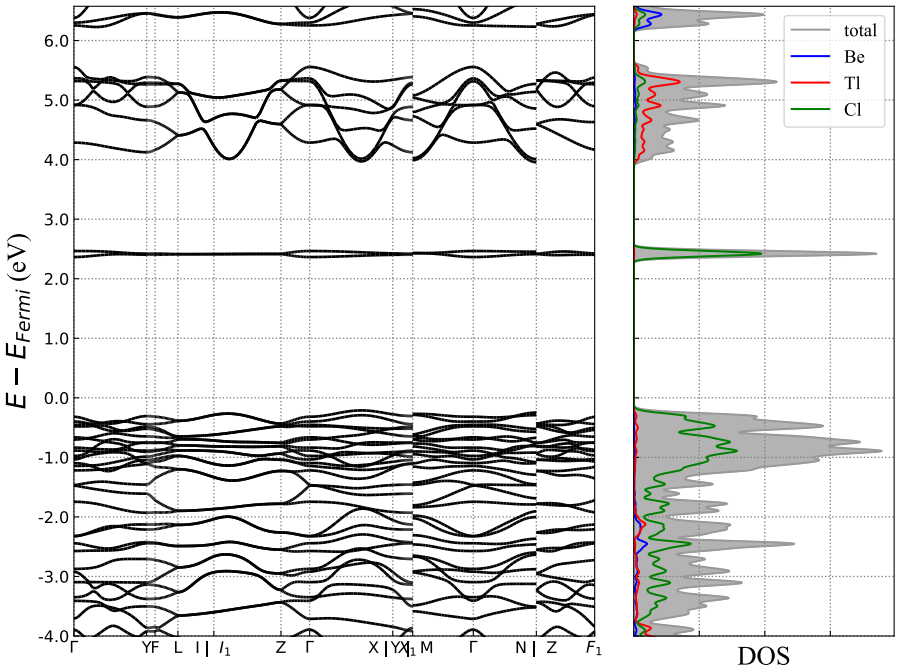
Fonte: elaborado pelo autor.

Figura B31: Estrutura de bandas e DOS do BeTlBr5.
Bandgap indireto de $E_g = 1.52$ eV.



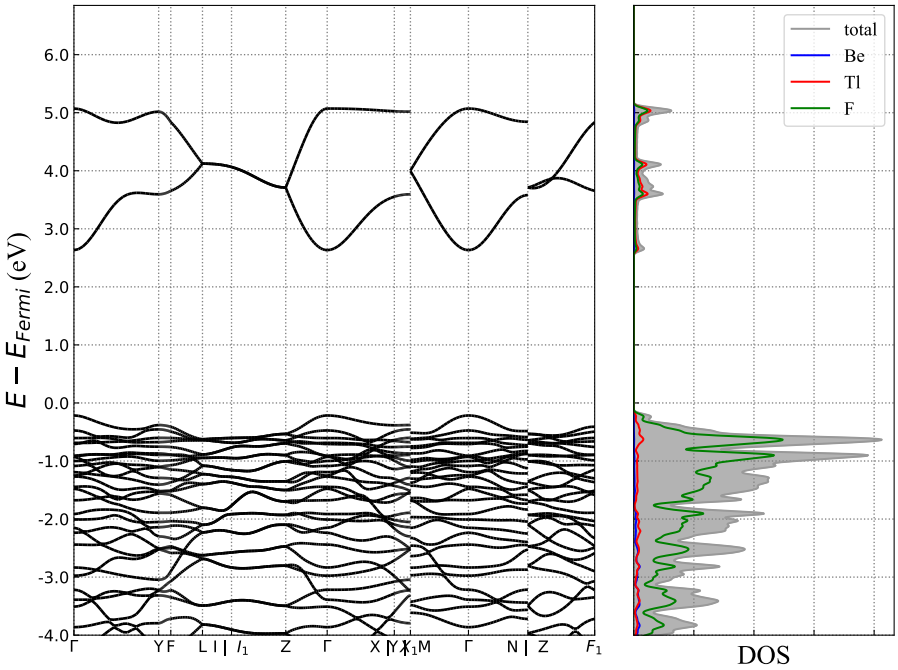
Fonte: elaborado pelo autor.

Figura B32: Estrutura de bandas e DOS do BeTlCl5.
Bandgap indireto de $E_g = 2.58$ eV.



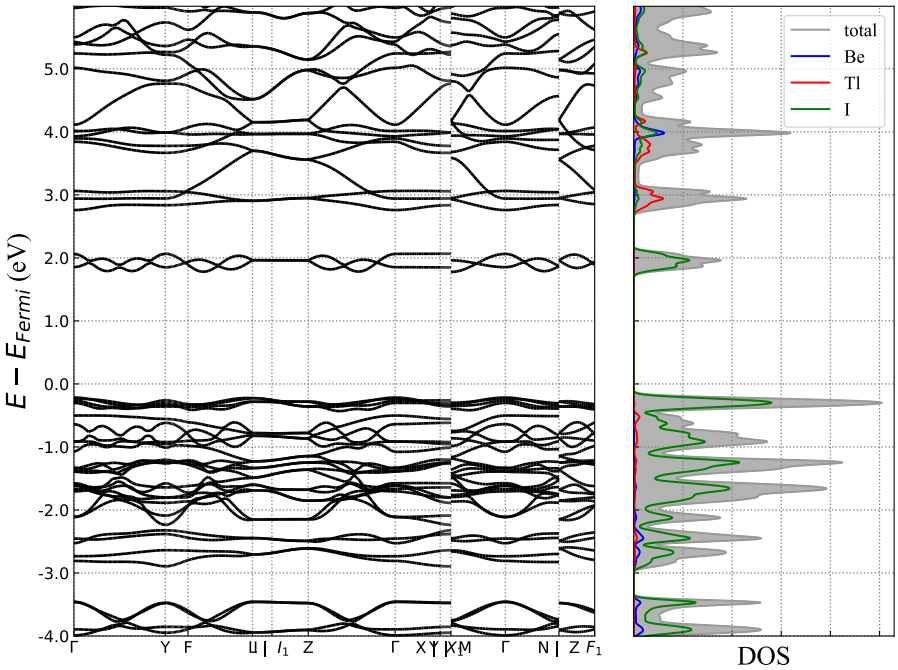
Fonte: elaborado pelo autor.

Figura B33: Estrutura de bandas e DOS do BeTlF5.
Bandgap direto de $E_g = 2.85$ eV.



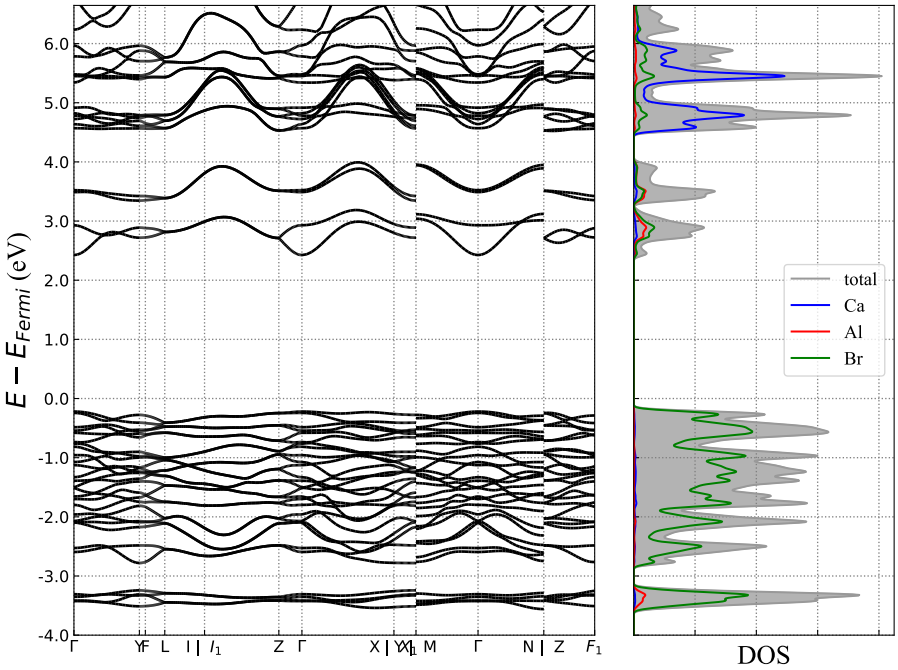
Fonte: elaborado pelo autor.

Figura B34: Estrutura de bandas e DOS do BeTlI5.
Bandgap indireto de $E_g = 1.99$ eV.



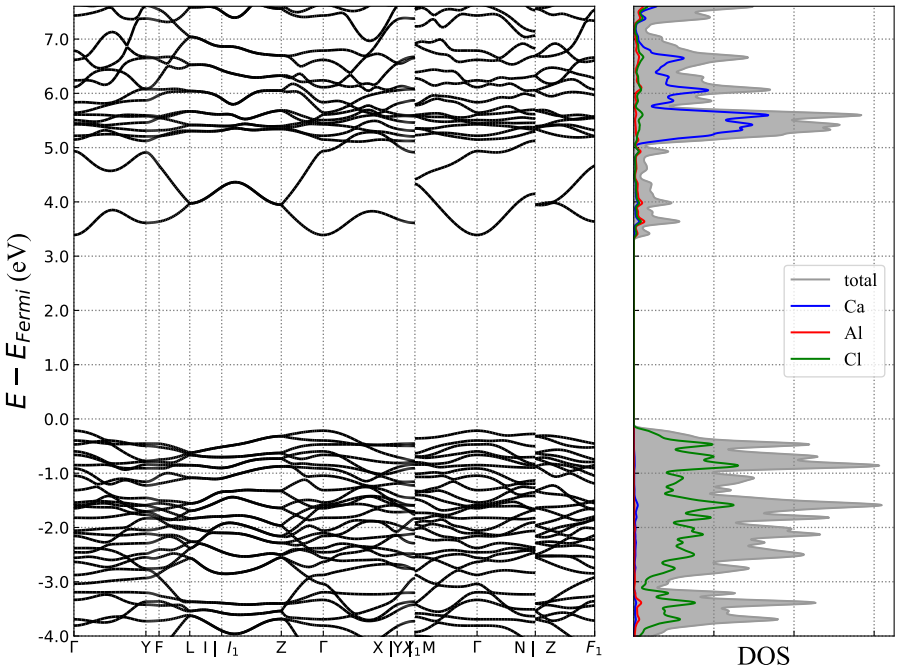
Fonte: elaborado pelo autor.

Figura B35: Estrutura de bandas e DOS do CaAlBr5.
Bandgap direto de $E_g = 2.65$ eV.



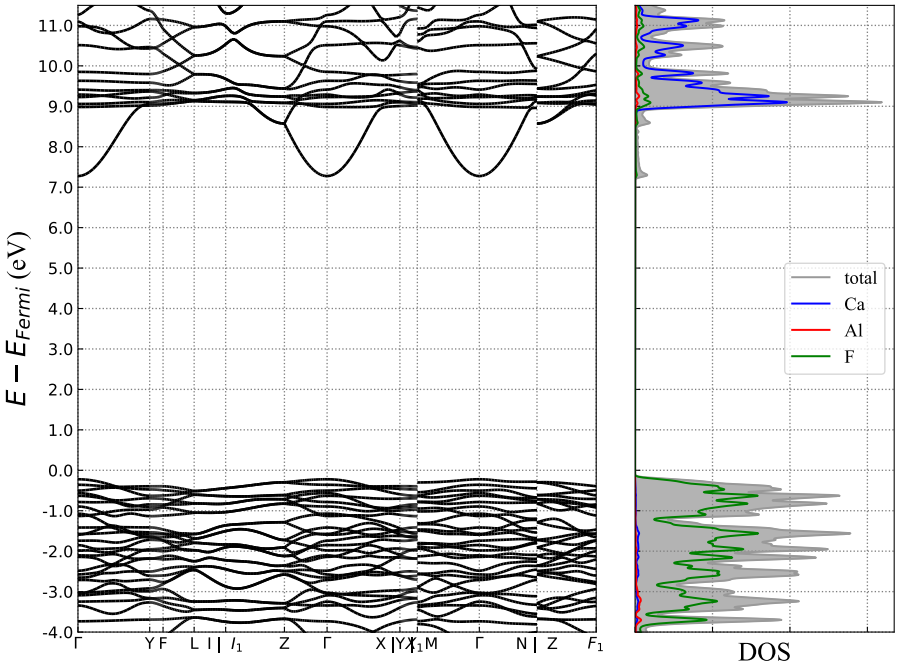
Fonte: elaborado pelo autor.

Figura B36: Estrutura de bandas e DOS do CaAlCl5.
Bandgap direto de $E_g = 3.6$ eV.



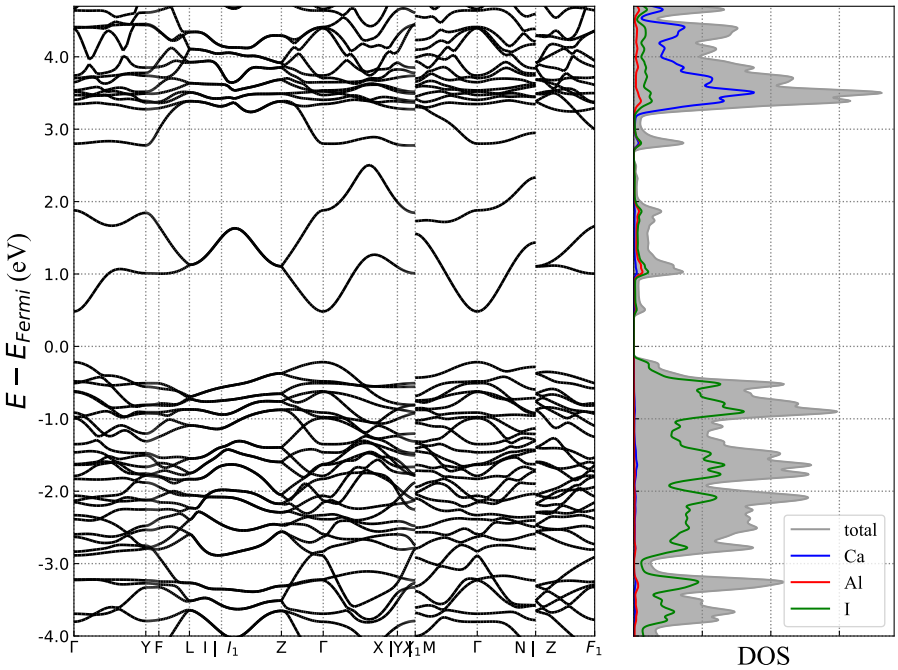
Fonte: elaborado pelo autor.

Figura B37: Estrutura de bandas e DOS do CaAlF5.
Bandgap direto de $E_g = 7.5$ eV.



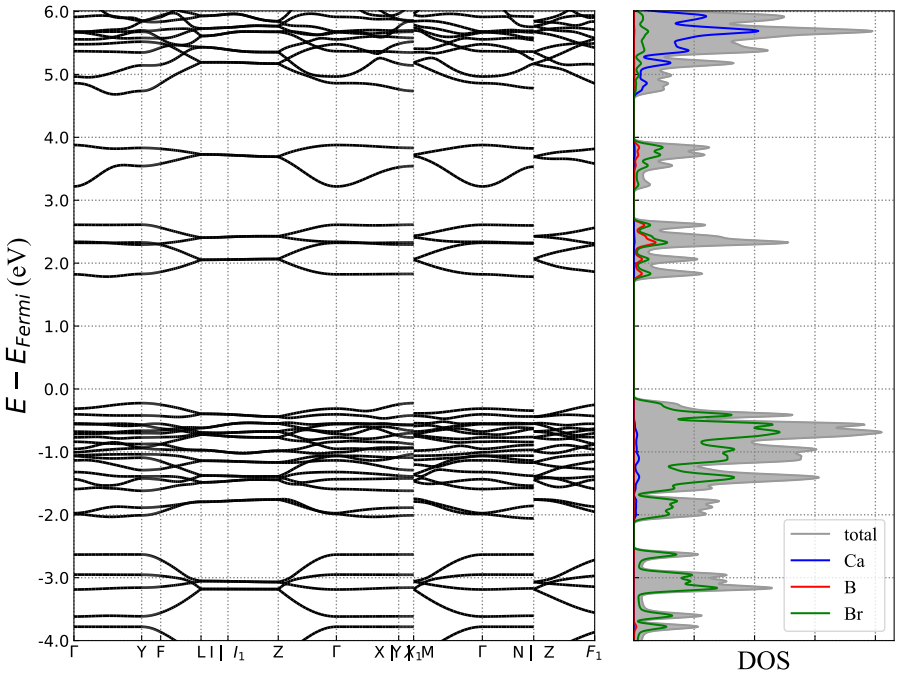
Fonte: elaborado pelo autor.

Figura B38: Estrutura de bandas e DOS do CaAlI5.
Bandgap direto de $E_g = 0.7$ eV.



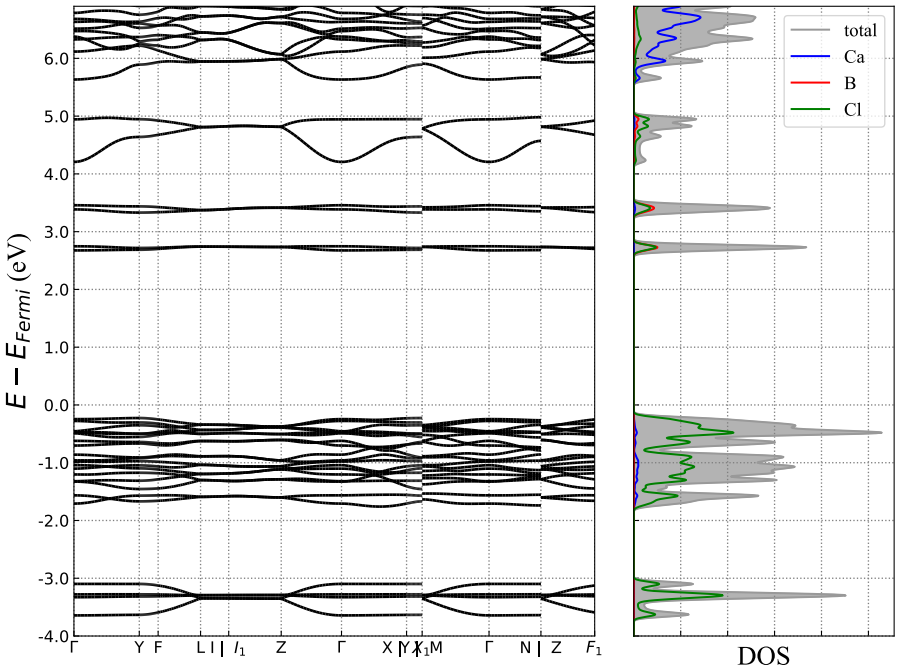
Fonte: elaborado pelo autor.

Figura B39: Estrutura de bandas e DOS do CaBBr5.
Bandgap indireto de $E_g = 2.01$ eV.



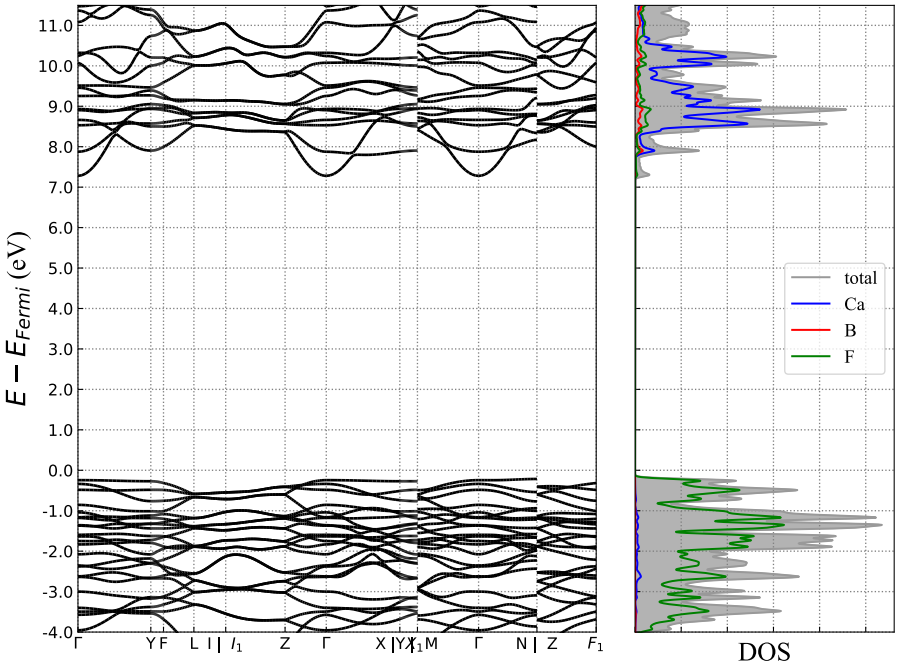
Fonte: elaborado pelo autor.

Figura B40: Estrutura de bandas e DOS do CaBCl5.
Bandgap indireto de $E_g = 2.9$ eV.



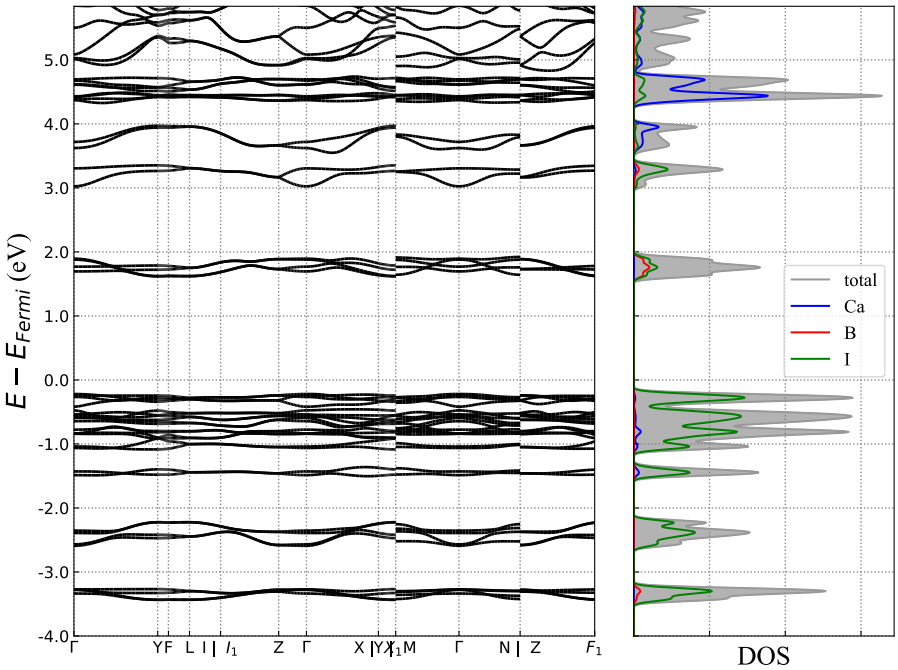
Fonte: elaborado pelo autor.

Figura B41: Estrutura de bandas e DOS do CaBF5.
Bandgap indireto de $E_g = 7.49$ eV.



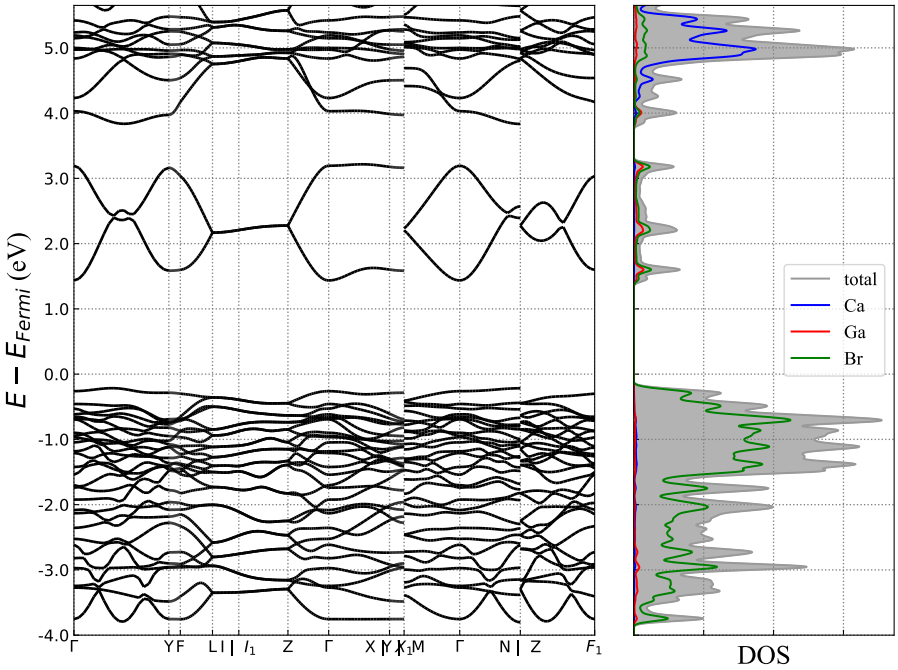
Fonte: elaborado pelo autor.

Figura B42: Estrutura de bandas e DOS do CaBI5.
Bandgap indireto de $E_g = 1.84$ eV.



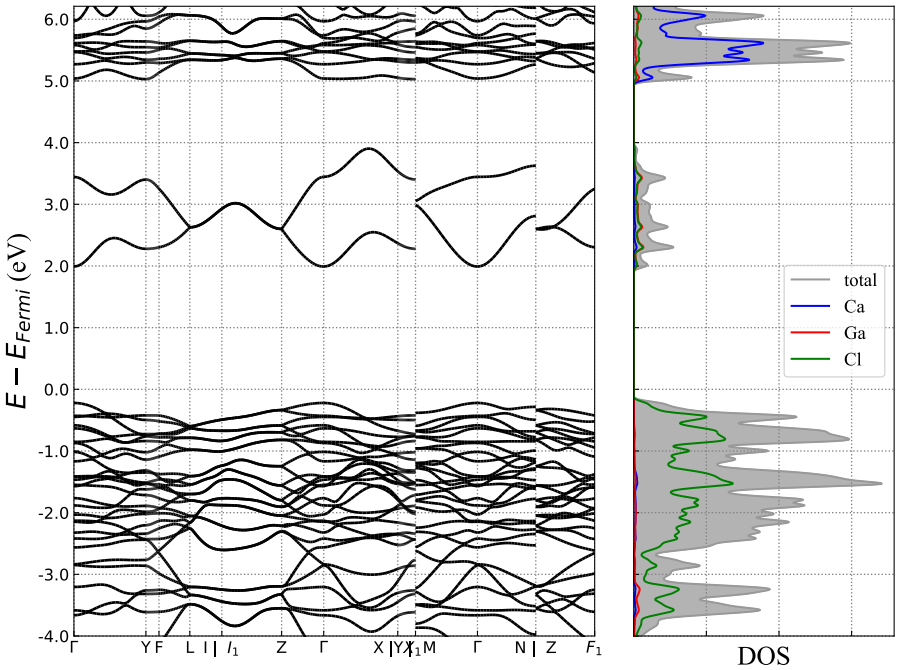
Fonte: elaborado pelo autor.

Figura B43: Estrutura de bandas e DOS do CaGaBr5.
Bandgap indireto de $E_g = 1.65$ eV.



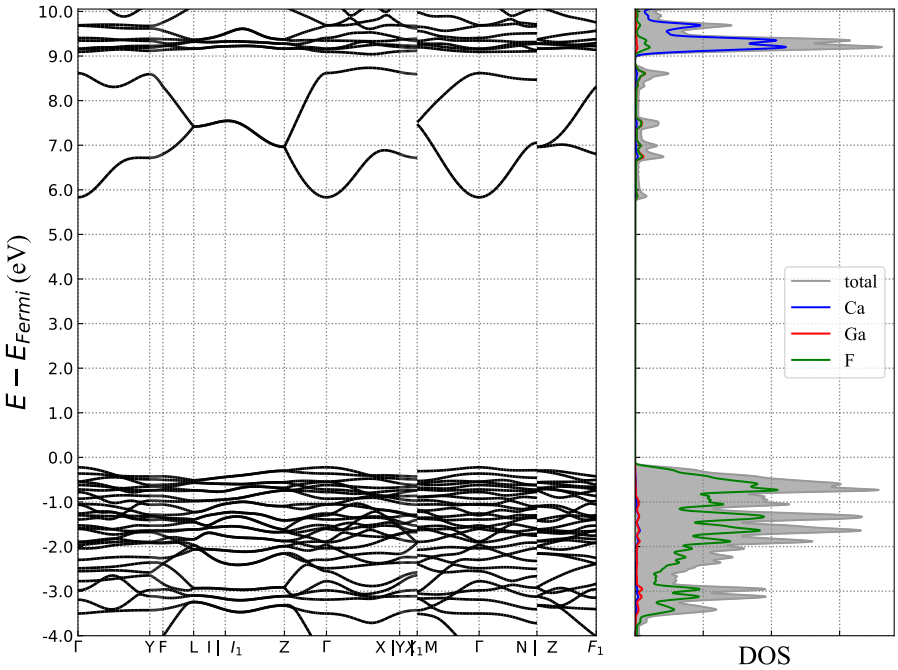
Fonte: elaborado pelo autor.

Figura B44: Estrutura de bandas e DOS do CaGaCl5.
Bandgap direto de $E_g = 2.21$ eV.



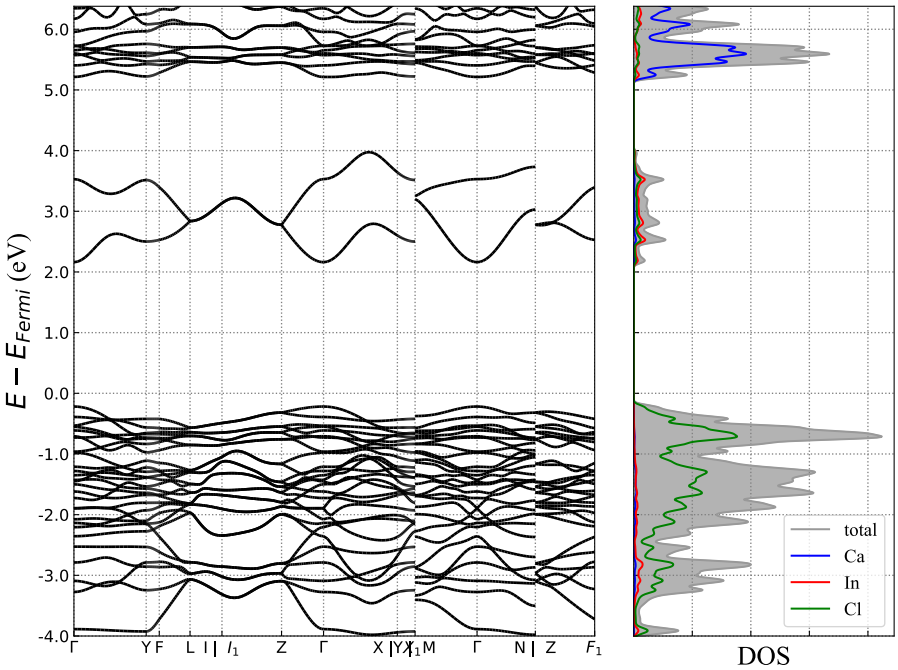
Fonte: elaborado pelo autor.

Figura B45: Estrutura de bandas e DOS do CaGaF5.
Bandgap direto de $E_g = 6.05$ eV.



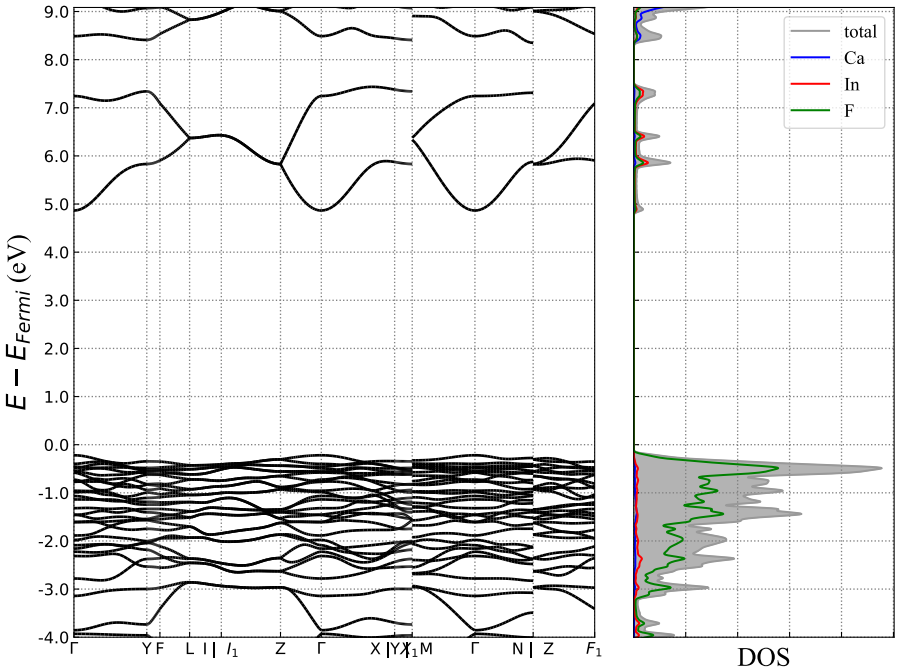
Fonte: elaborado pelo autor.

Figura B46: Estrutura de bandas e DOS do CaInCl5.
Bandgap direto de $E_g = 2.38$ eV.



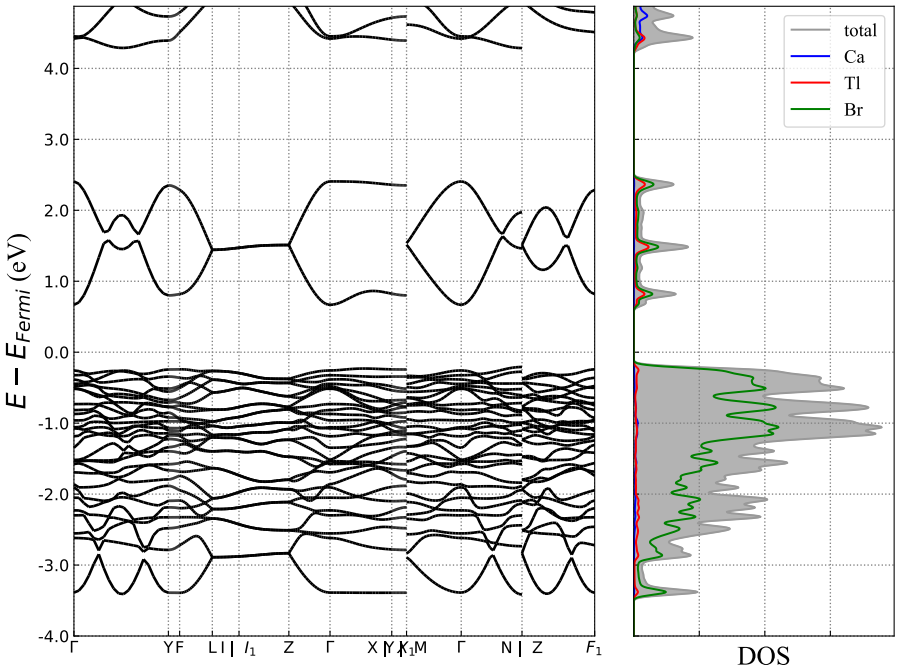
Fonte: elaborado pelo autor.

Figura B47: Estrutura de bandas e DOS do CaInF5.
Bandgap direto de $E_g = 5.08$ eV.



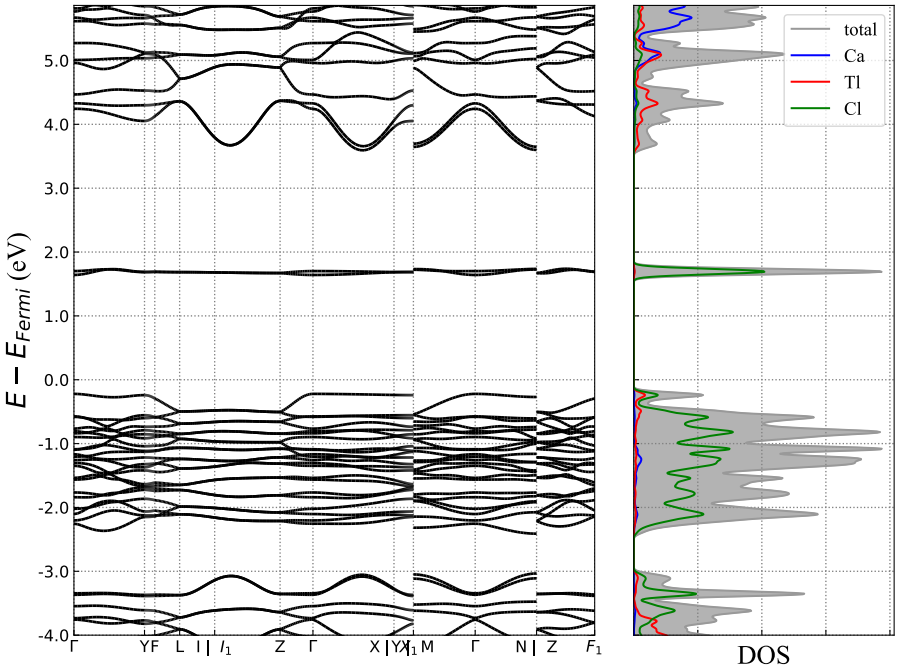
Fonte: elaborado pelo autor.

Figura B48: Estrutura de bandas e DOS do CaTlBr5.
Bandgap indireto de $E_g = 0.88$ eV.



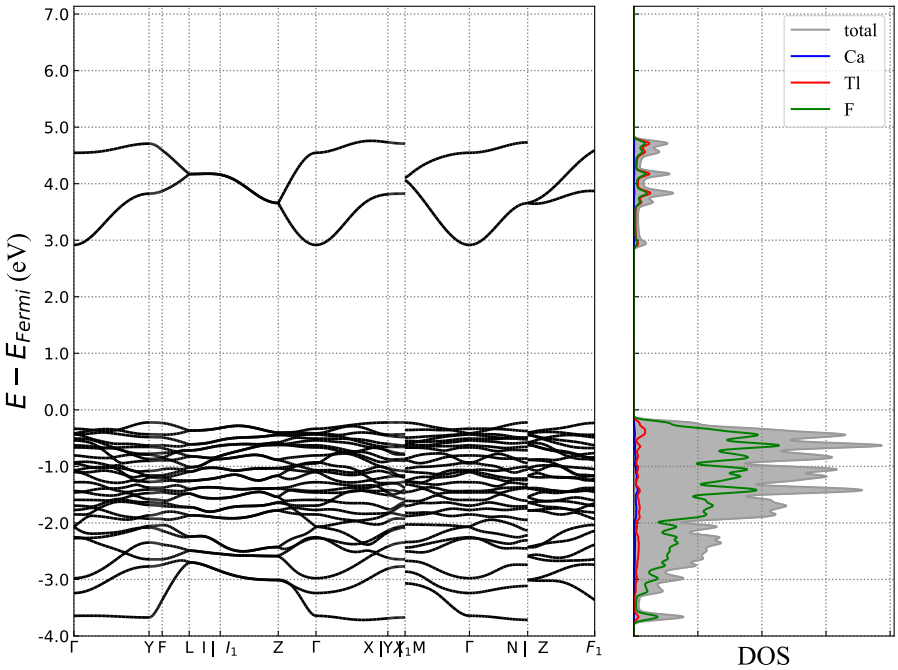
Fonte: elaborado pelo autor.

Figura B49: Estrutura de bandas e DOS do CaTlCl5.
Bandgap direto de $E_g = 1.86$ eV.



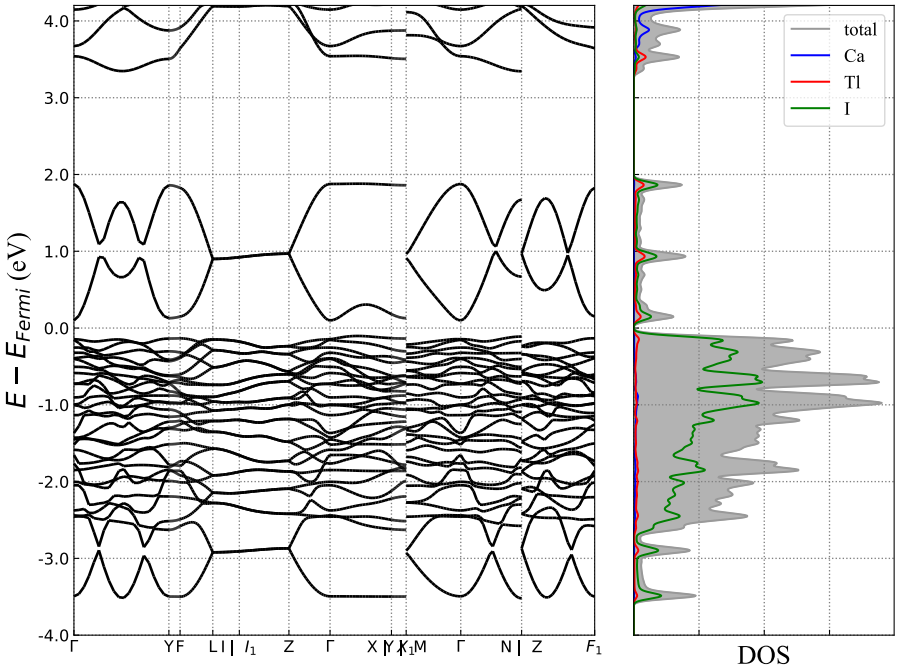
Fonte: elaborado pelo autor.

Figura B50: Estrutura de bandas e DOS do CaTlF5.
Bandgap indireto de $E_g = 3.14$ eV.



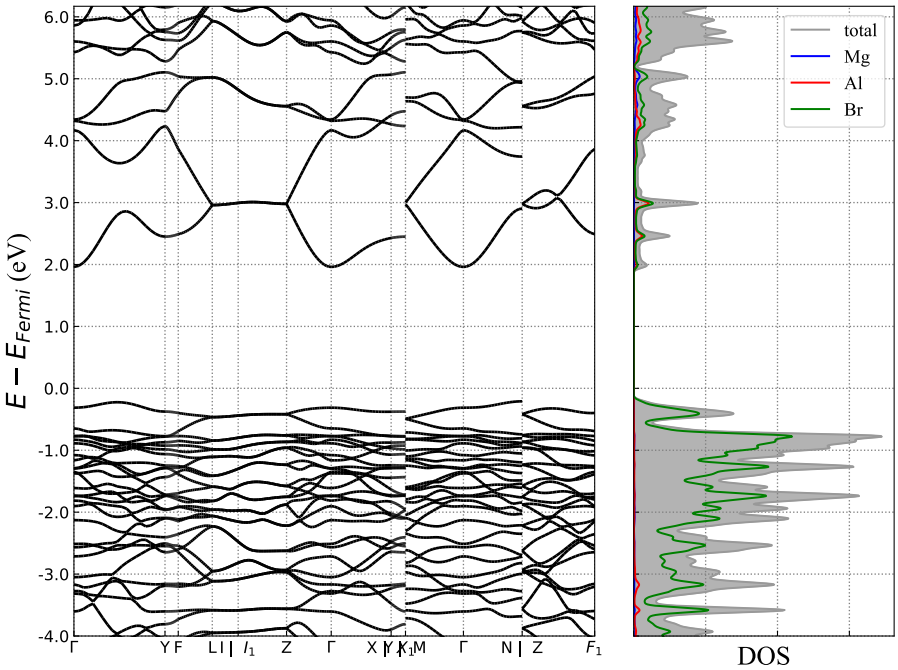
Fonte: elaborado pelo autor.

Figura B51: Estrutura de bandas e DOS do CaTlI5.
Bandgap indireto de $E_g = 0.2$ eV.



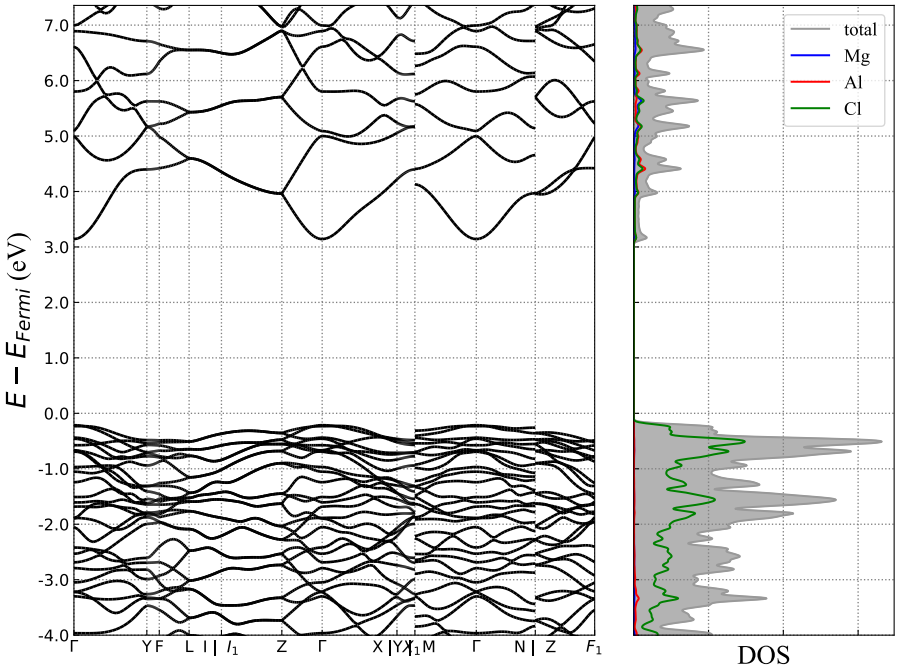
Fonte: elaborado pelo autor.

Figura B52: Estrutura de bandas e DOS do MgAlBr5.
Bandgap indireto de $E_g = 2.17$ eV.



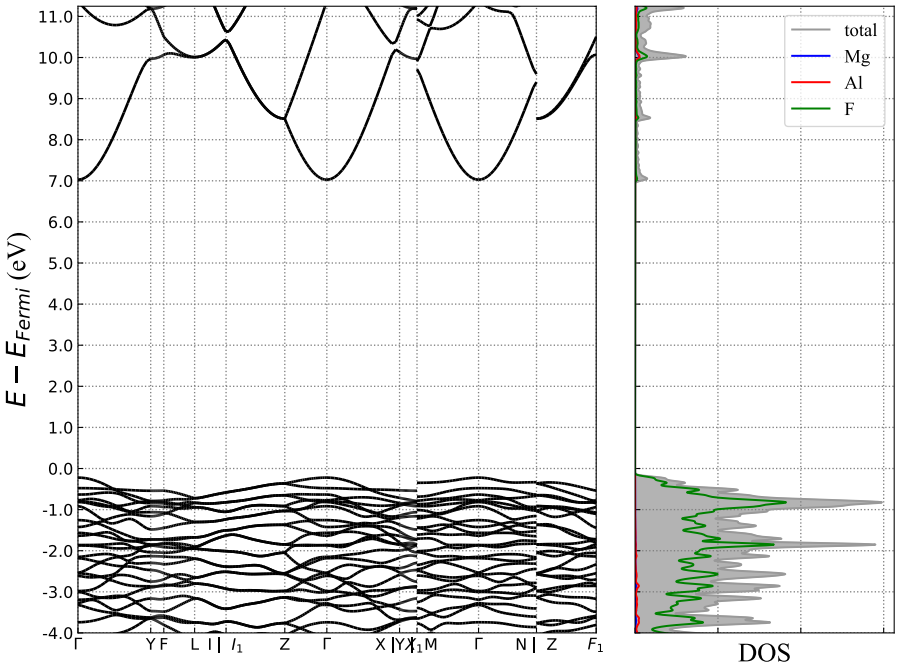
Fonte: elaborado pelo autor.

Figura B53: Estrutura de bandas e DOS do MgAlCl5.
Bandgap direto de $E_g = 3.36$ eV.



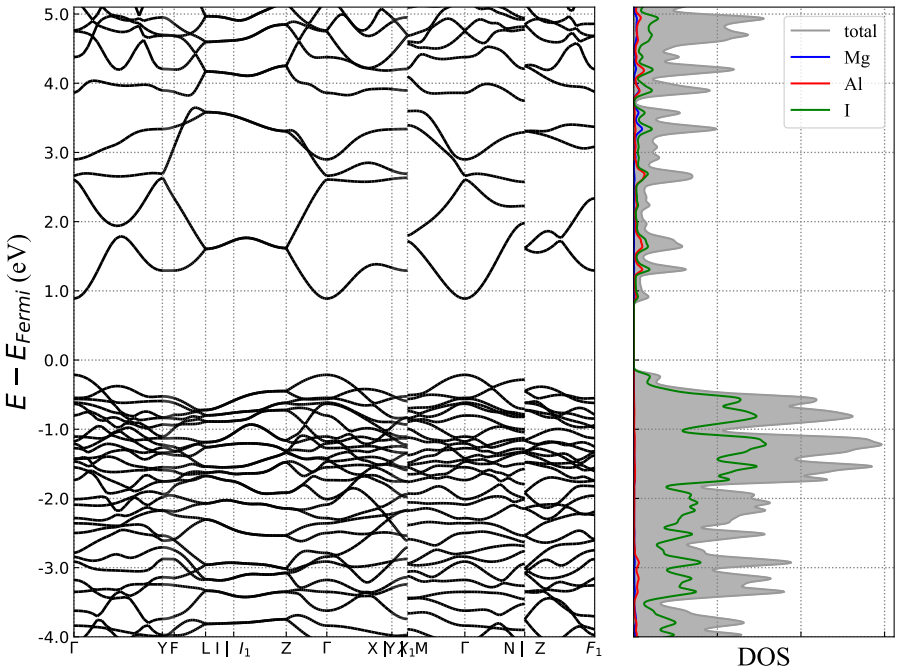
Fonte: elaborado pelo autor.

Figura B54: Estrutura de bandas e DOS do MgAlF5.
Bandgap direto de $E_g = 7.25$ eV.



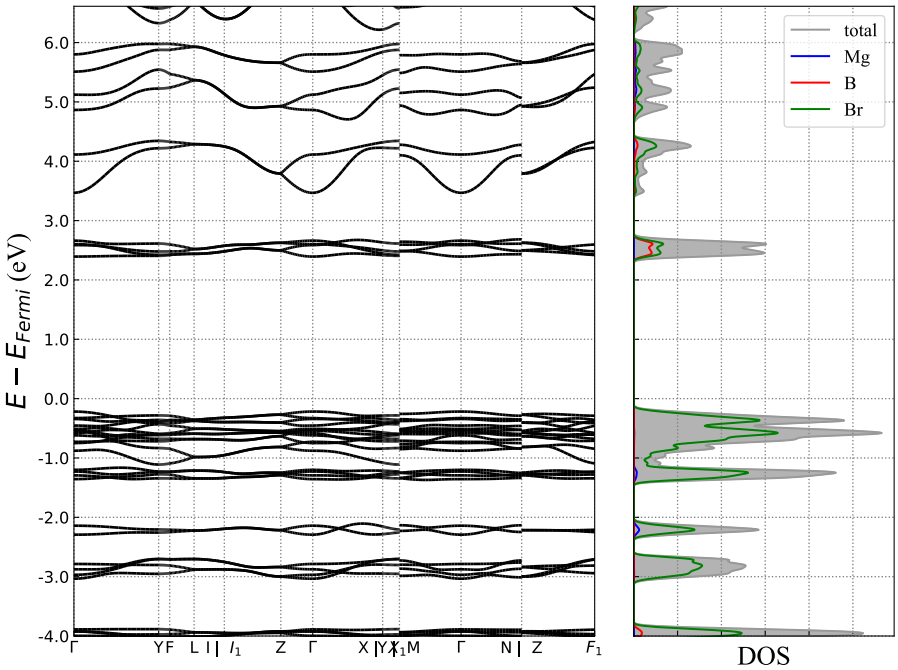
Fonte: elaborado pelo autor.

Figura B55: Estrutura de bandas e DOS do MgAlI5.
Bandgap direto de $E_g = 1.1$ eV.



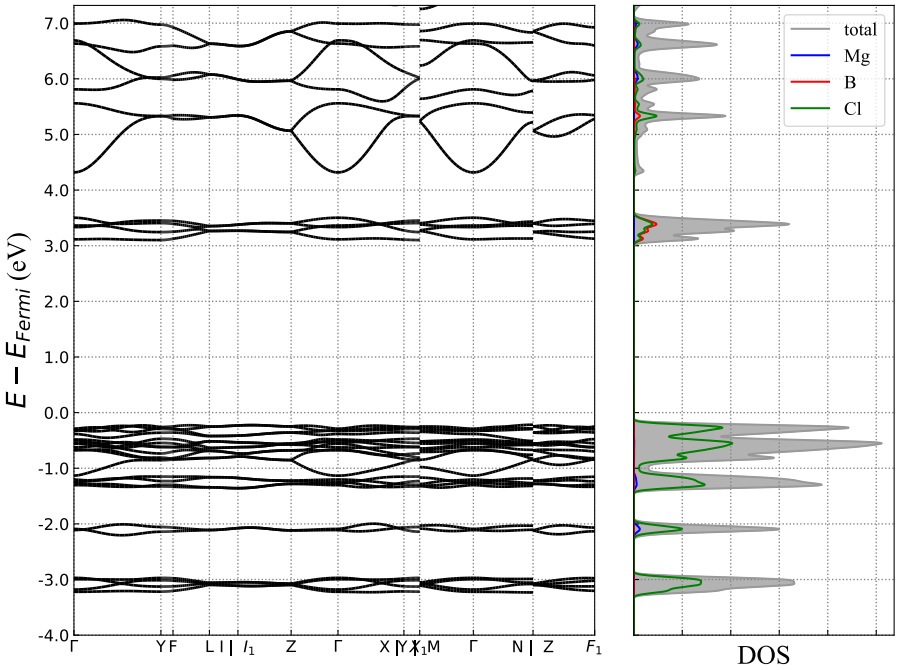
Fonte: elaborado pelo autor.

Figura B56: Estrutura de bandas e DOS do MgBBr5.
Bandgap direto de $E_g = 2.61$ eV.



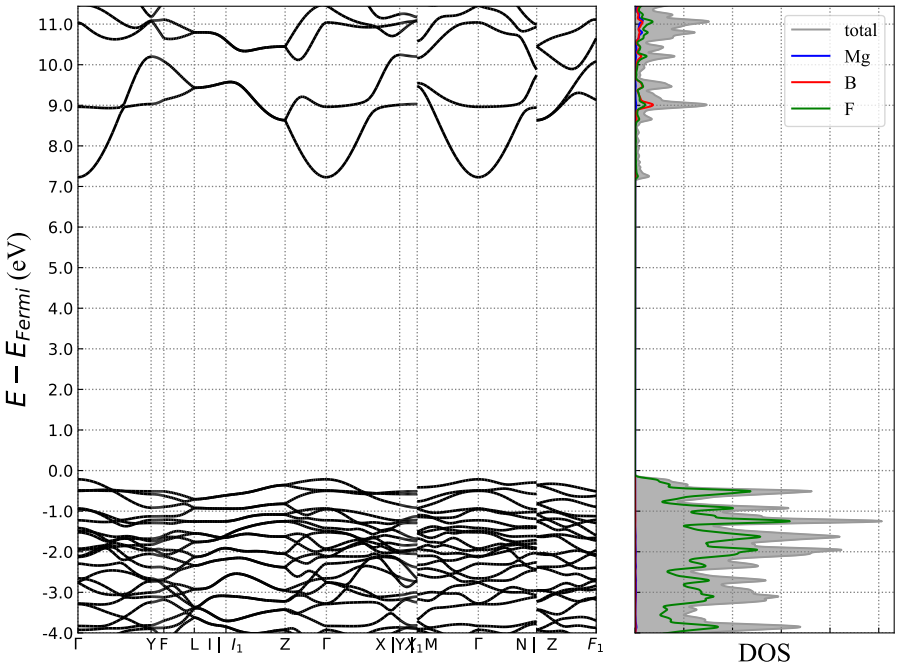
Fonte: elaborado pelo autor.

Figura B57: Estrutura de bandas e DOS do MgBCl5.
Bandgap indireto de $E_g = 3.32$ eV.



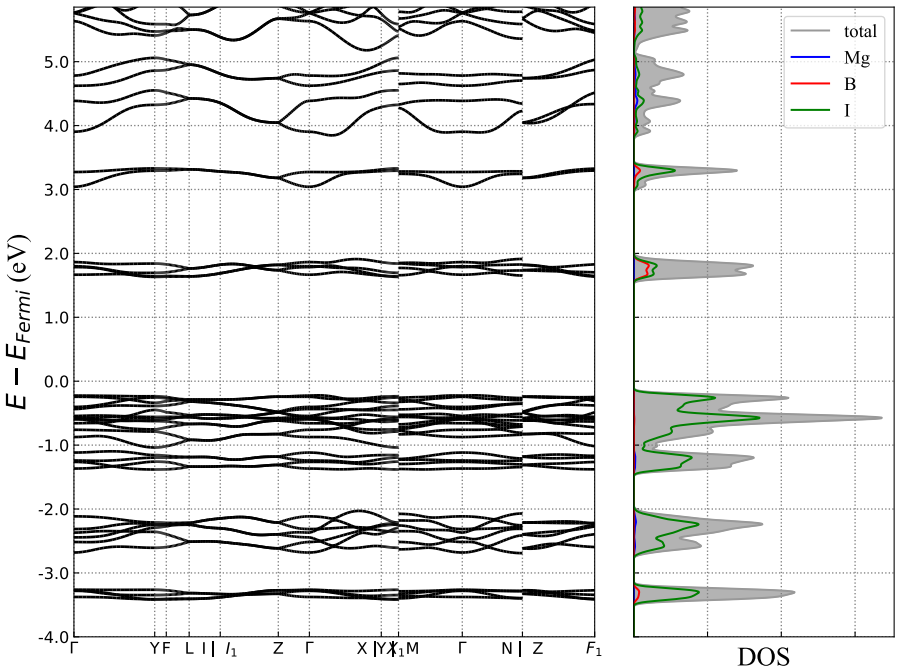
Fonte: elaborado pelo autor.

Figura B58: Estrutura de bandas e DOS do MgBF5.
Bandgap direto de $E_g = 7.44$ eV.



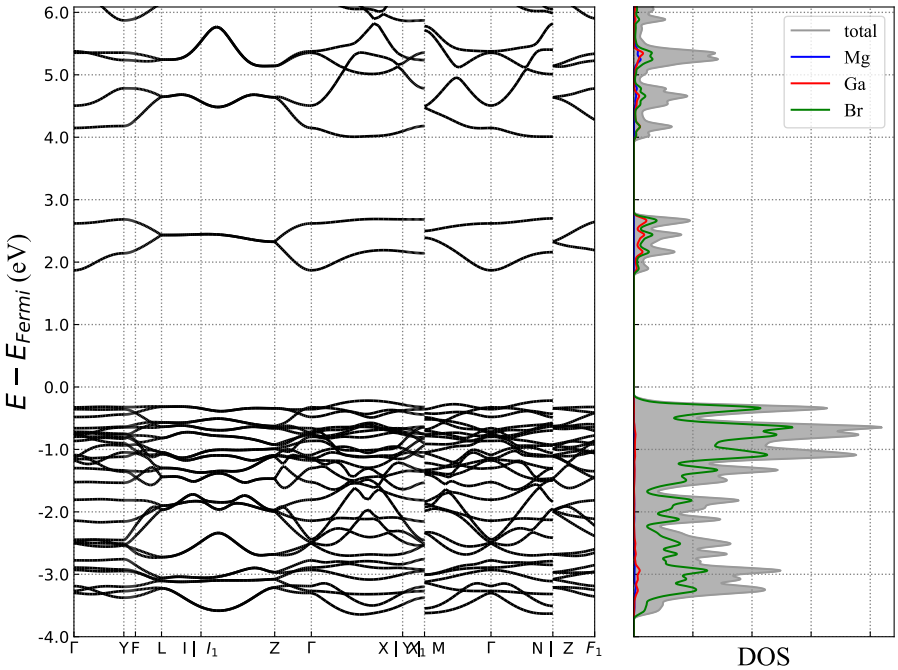
Fonte: elaborado pelo autor.

Figura B59: Estrutura de bandas e DOS do MgBI5.
Bandgap indireto de $E_g = 1.85$ eV.



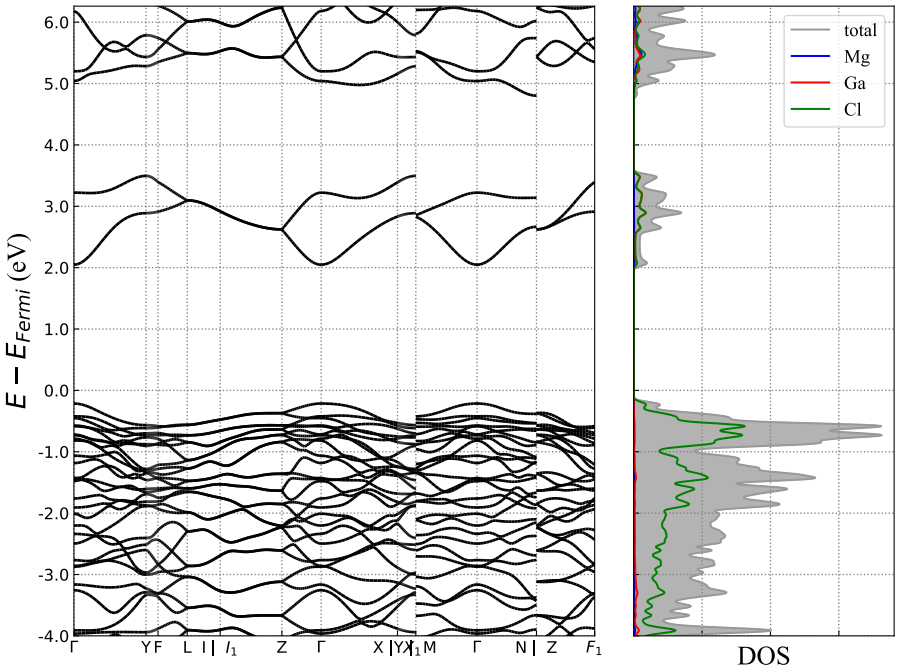
Fonte: elaborado pelo autor.

Figura B60: Estrutura de bandas e DOS do MgGaBr5.
Bandgap indireto de $E_g = 2.09$ eV.



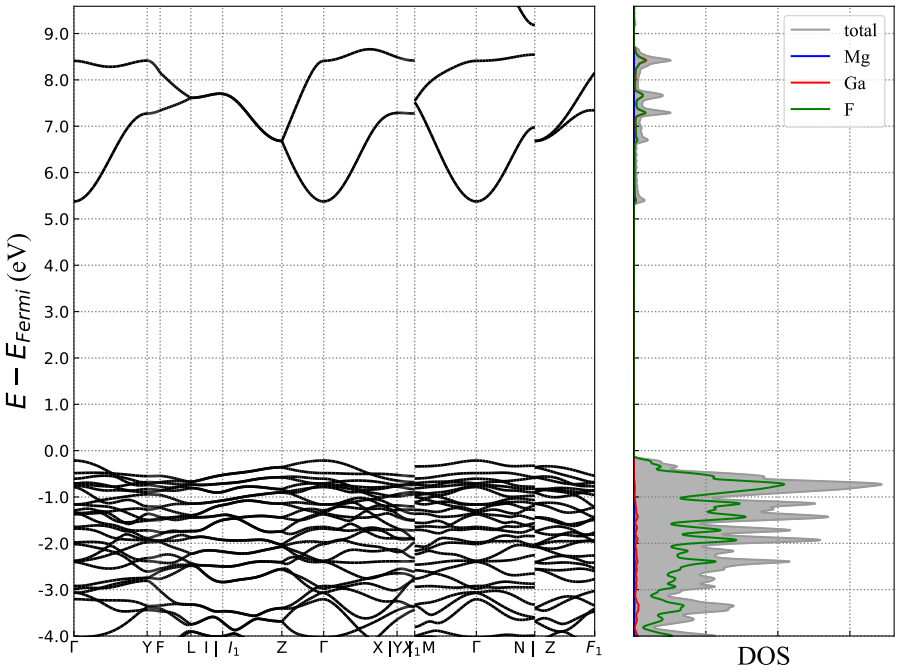
Fonte: elaborado pelo autor.

Figura B61: Estrutura de bandas e DOS do MgGaCl5.
Bandgap direto de $E_g = 2.26$ eV.



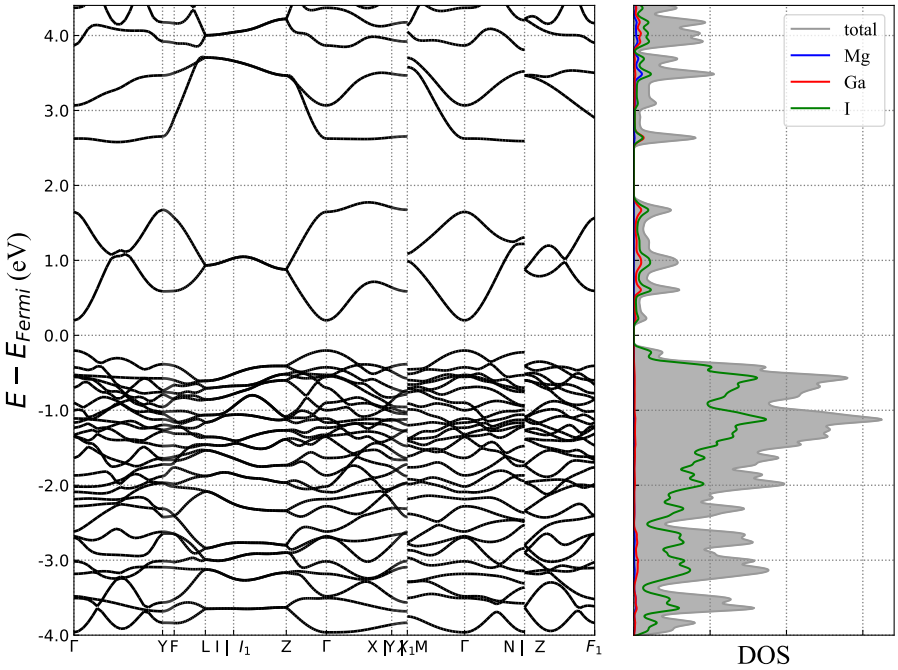
Fonte: elaborado pelo autor.

Figura B62: Estrutura de bandas e DOS do MgGaF5.
Bandgap direto de $E_g = 5.59$ eV.



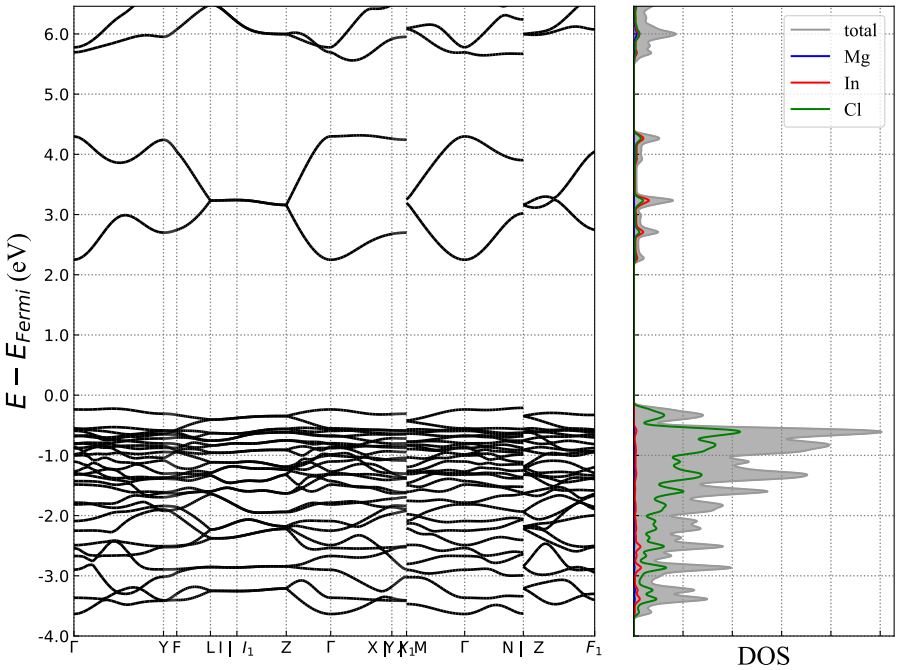
Fonte: elaborado pelo autor.

Figura B63: Estrutura de bandas e DOS do MgGaI5.
Bandgap direto de $E_g = 0.4$ eV.



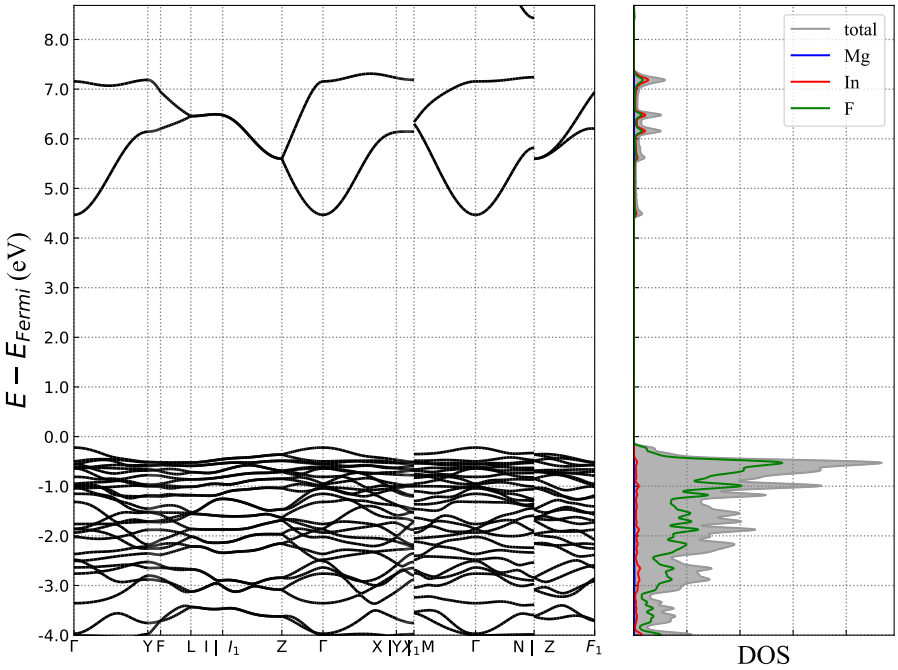
Fonte: elaborado pelo autor.

Figura B64: Estrutura de bandas e DOS do MgInCl5.
Bandgap indireto de $E_g = 2.46$ eV.



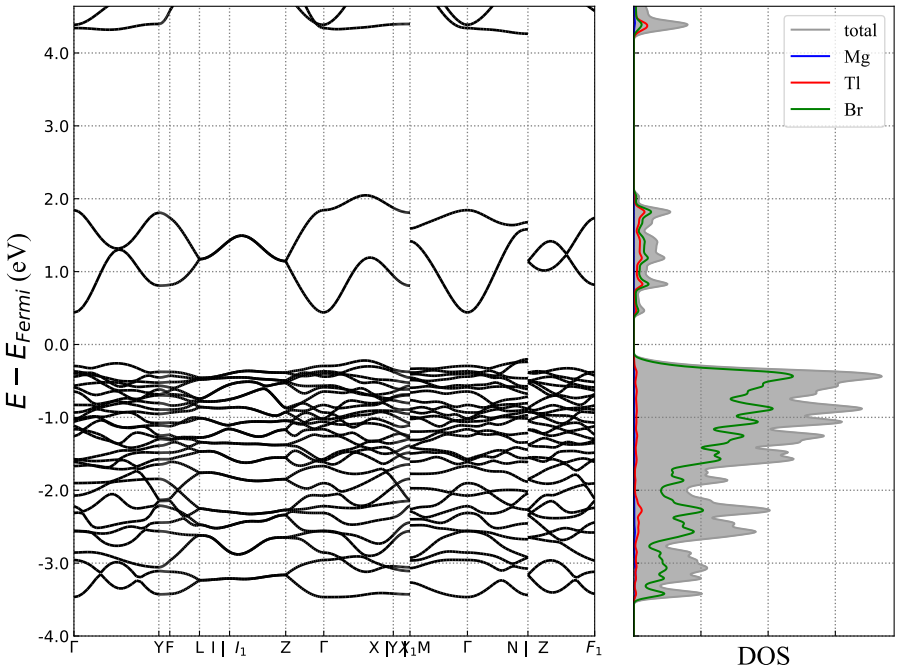
Fonte: elaborado pelo autor.

Figura B65: Estrutura de bandas e DOS do MgInF5.
Bandgap direto de $E_g = 4.69$ eV.



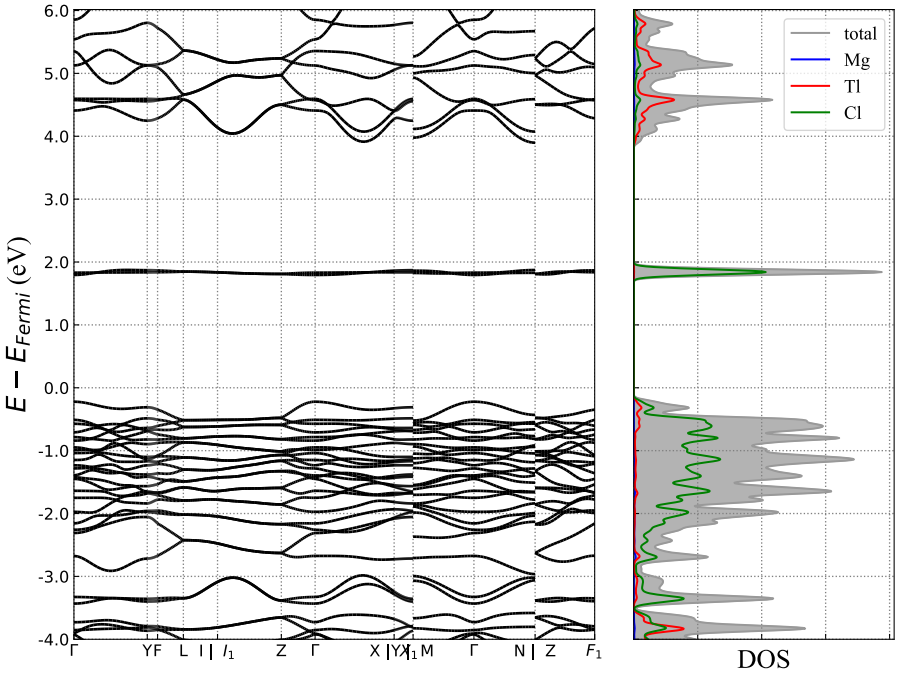
Fonte: elaborado pelo autor.

Figura B66: Estrutura de bandas e DOS do MgTlBr5.
Bandgap indireto de $E_g = 0.64$ eV.



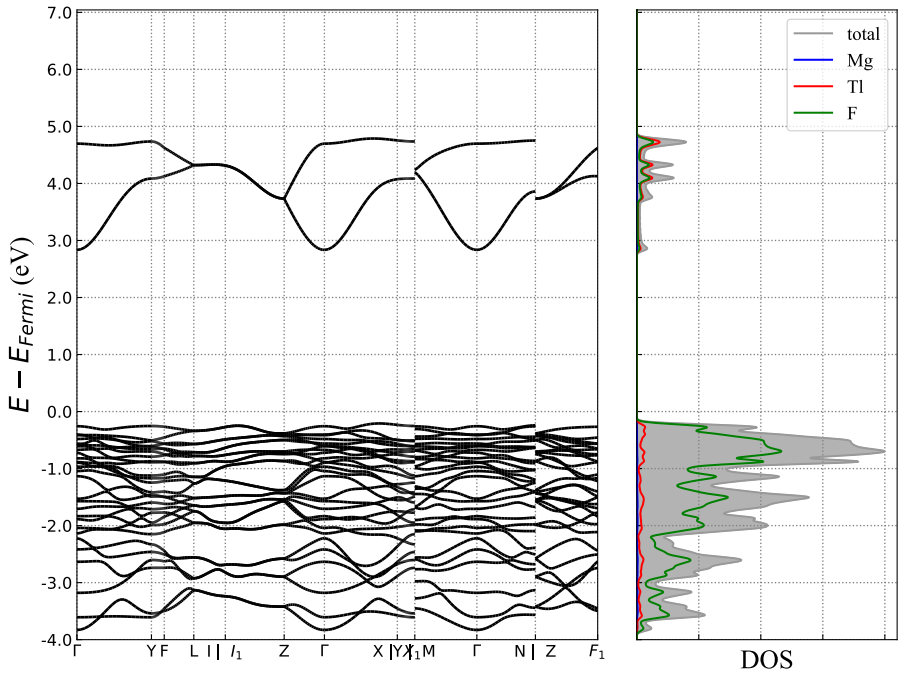
Fonte: elaborado pelo autor.

Figura B67: Estrutura de bandas e DOS do MgTlCl5.
Bandgap direto de $E_g = 2.02$ eV.



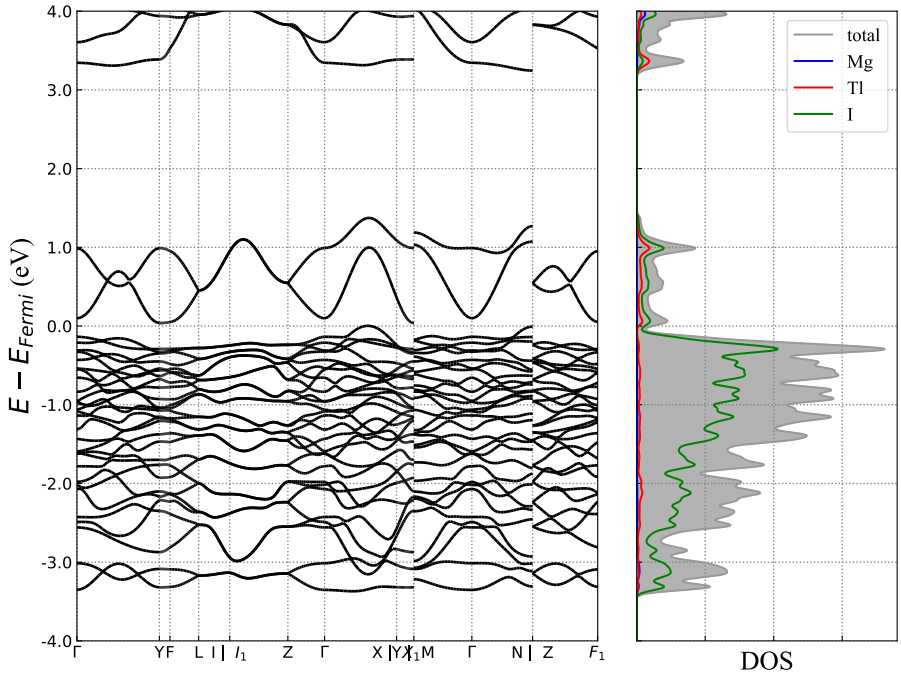
Fonte: elaborado pelo autor.

Figura B68: Estrutura de bandas e DOS do MgTlF5.
Bandgap indireto de $E_g = 3.04$ eV.



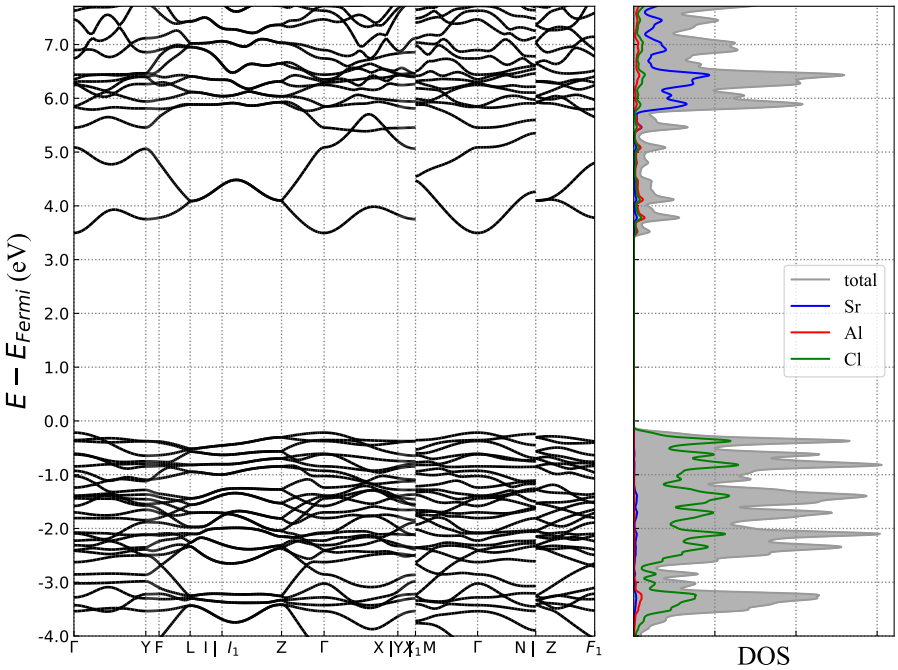
Fonte: elaborado pelo autor.

Figura B69: Estrutura de bandas e DOS do MgTlI5.
Bandgap indireto de $E_g = 0.0$ eV.



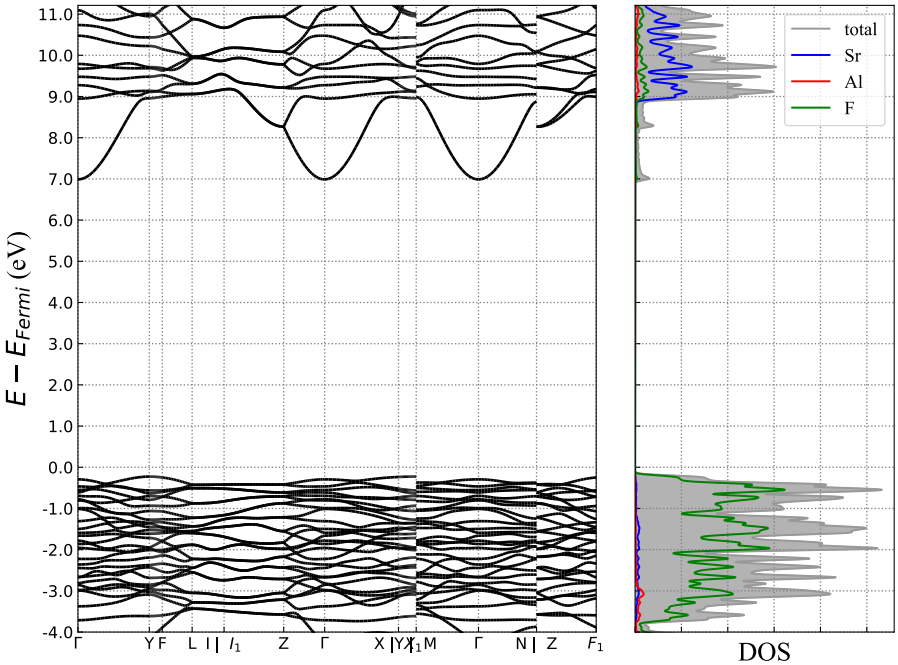
Fonte: elaborado pelo autor.

Figura B70: Estrutura de bandas e DOS do SrAlCl5.
Bandgap direto de $E_g = 3.71$ eV.



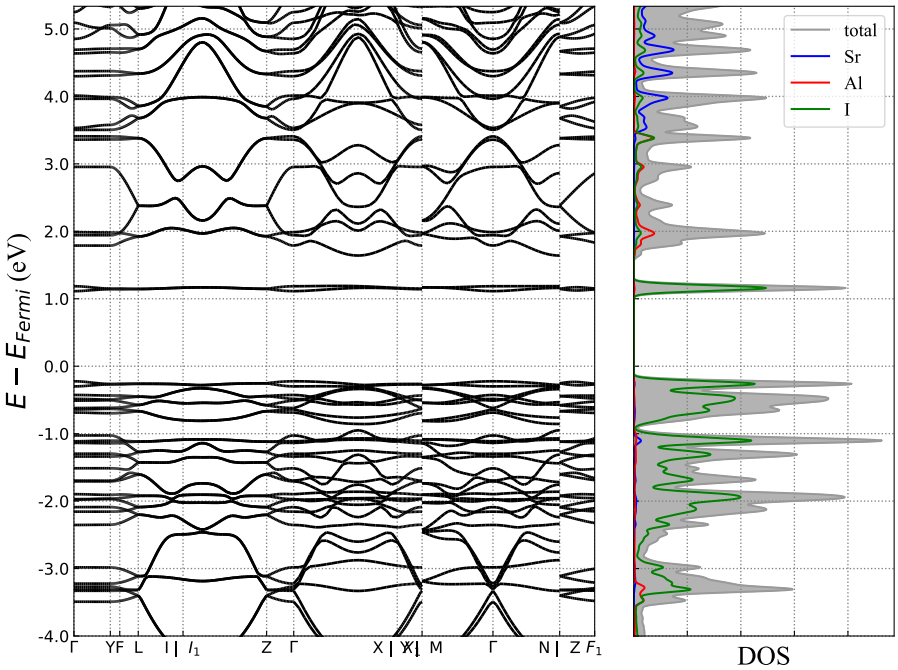
Fonte: elaborado pelo autor.

Figura B71: Estrutura de bandas e DOS do SrAlF5.
Bandgap indireto de $E_g = 7.21$ eV.



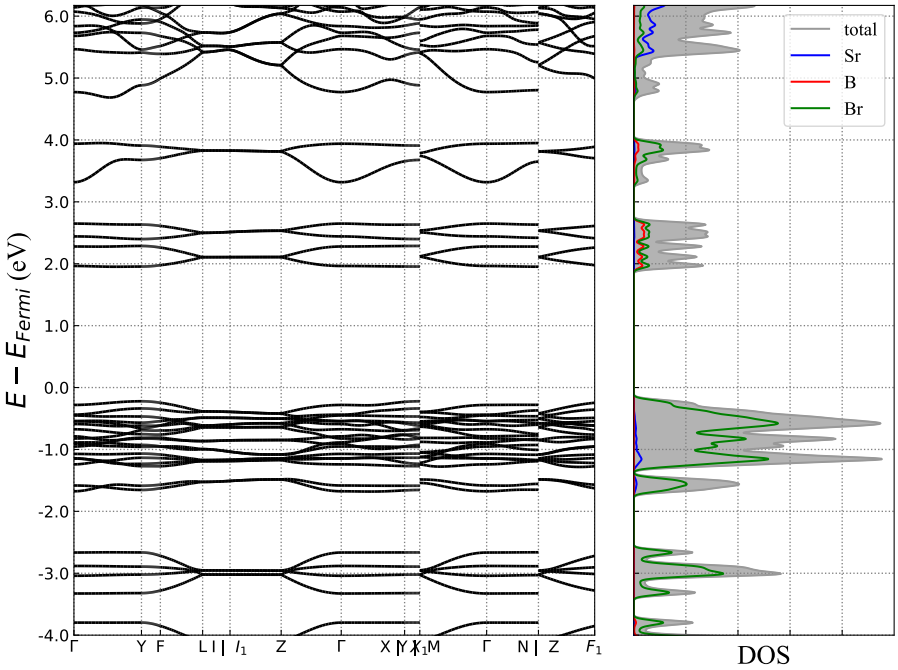
Fonte: elaborado pelo autor.

Figura B72: Estrutura de bandas e DOS do SrAlI5.
Bandgap direto de $E_g = 1.34$ eV.



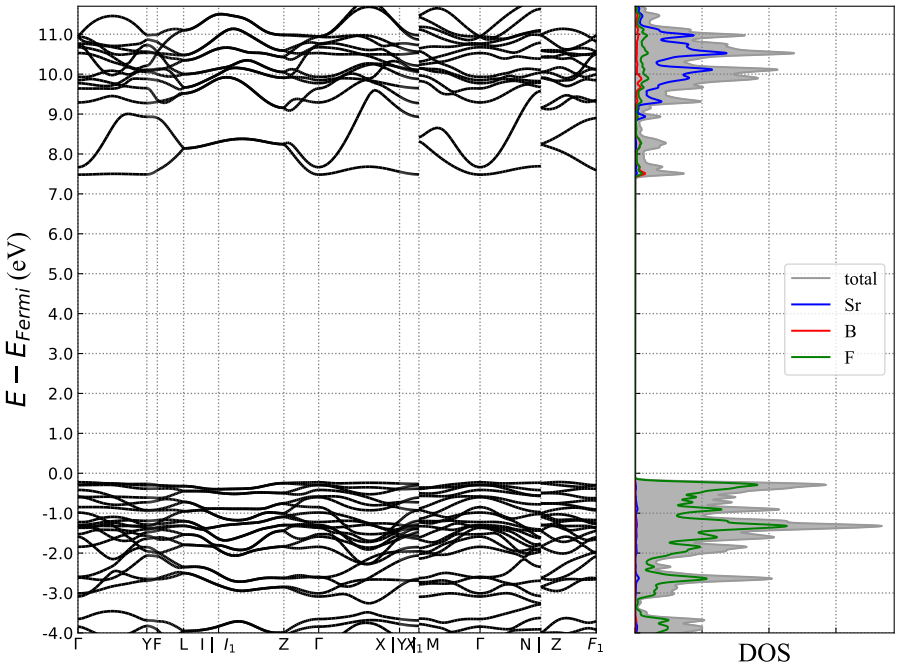
Fonte: elaborado pelo autor.

Figura B73: Estrutura de bandas e DOS do SrBBr5.
Bandgap indireto de $E_g = 2.17$ eV.



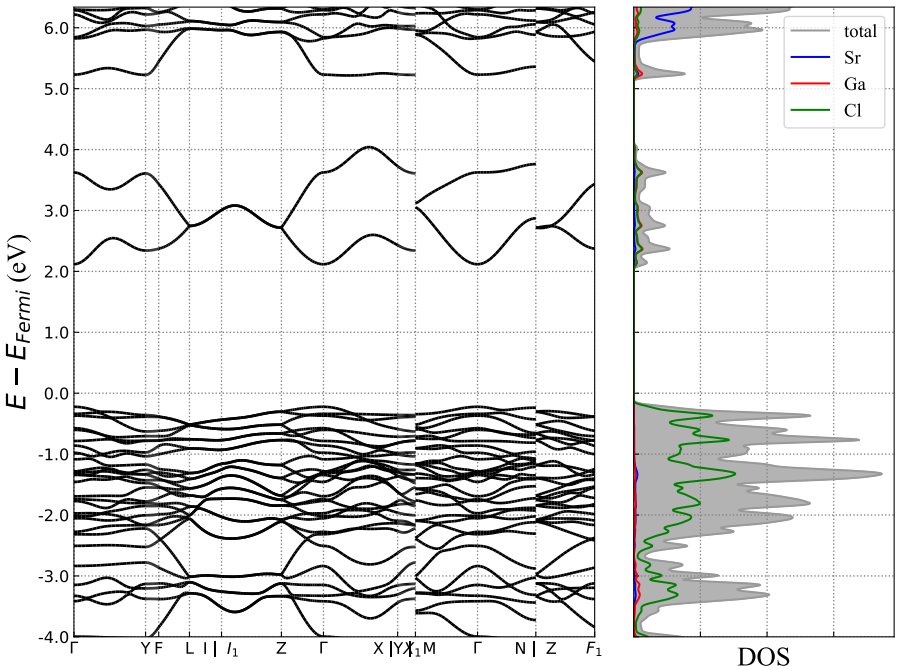
Fonte: elaborado pelo autor.

Figura B74: Estrutura de bandas e DOS do SrBF5.
Bandgap indireto de $E_g = 7.7$ eV.



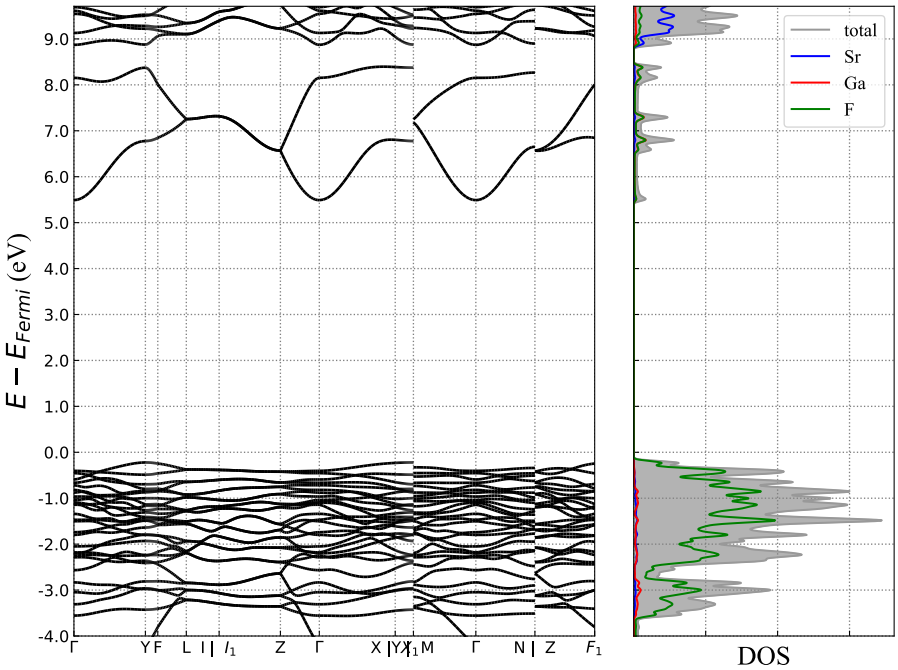
Fonte: elaborado pelo autor.

Figura B75: Estrutura de bandas e DOS do SrGaCl5.
Bandgap direto de $E_g = 2.34$ eV.



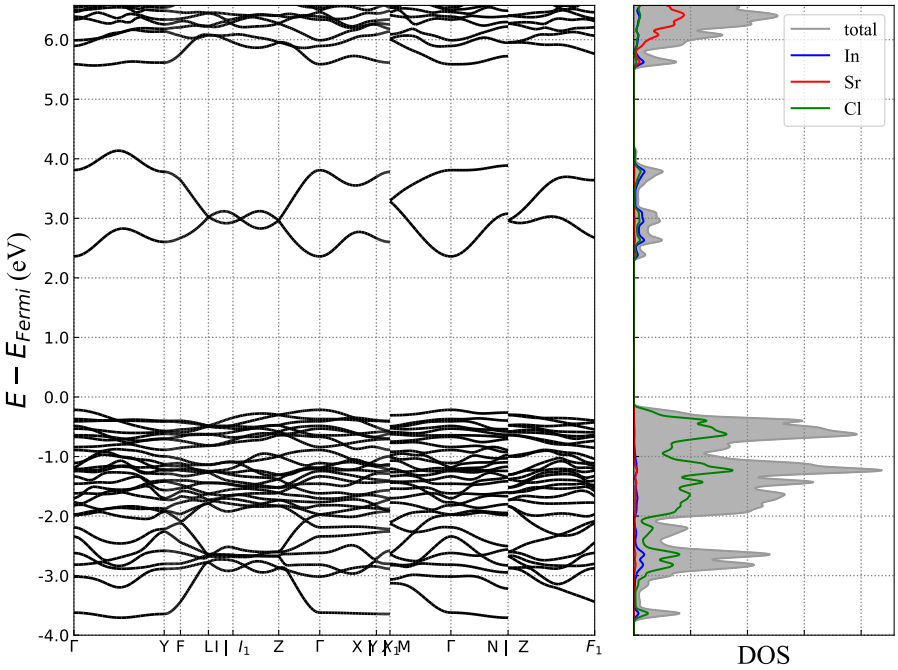
Fonte: elaborado pelo autor.

Figura B76: Estrutura de bandas e DOS do SrGaF5.
Bandgap indireto de $E_g = 5.71$ eV.



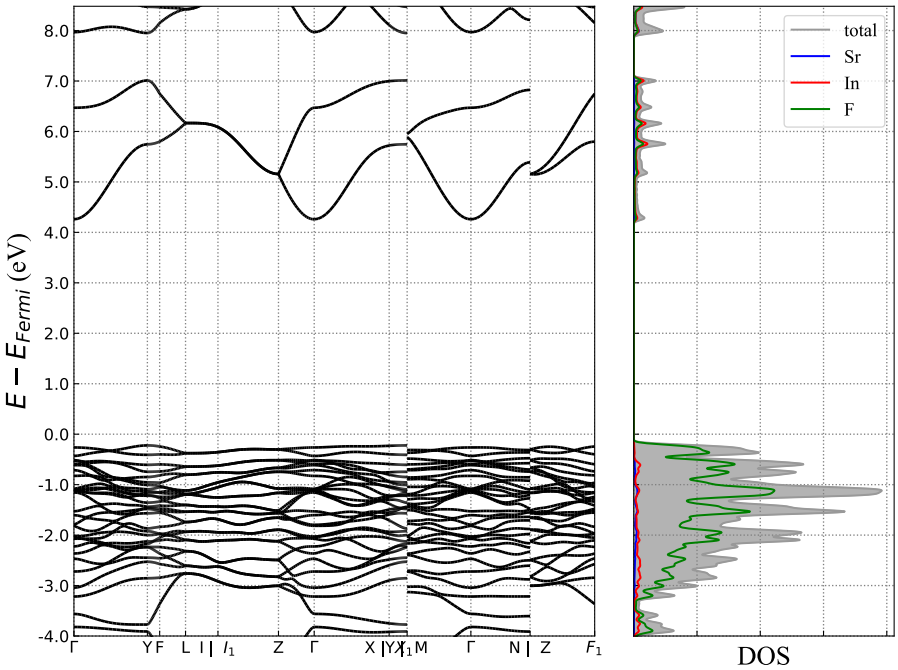
Fonte: elaborado pelo autor.

Figura B77: Estrutura de bandas e DOS do SrInCl5.
Bandgap direto de $E_g = 2.58$ eV.



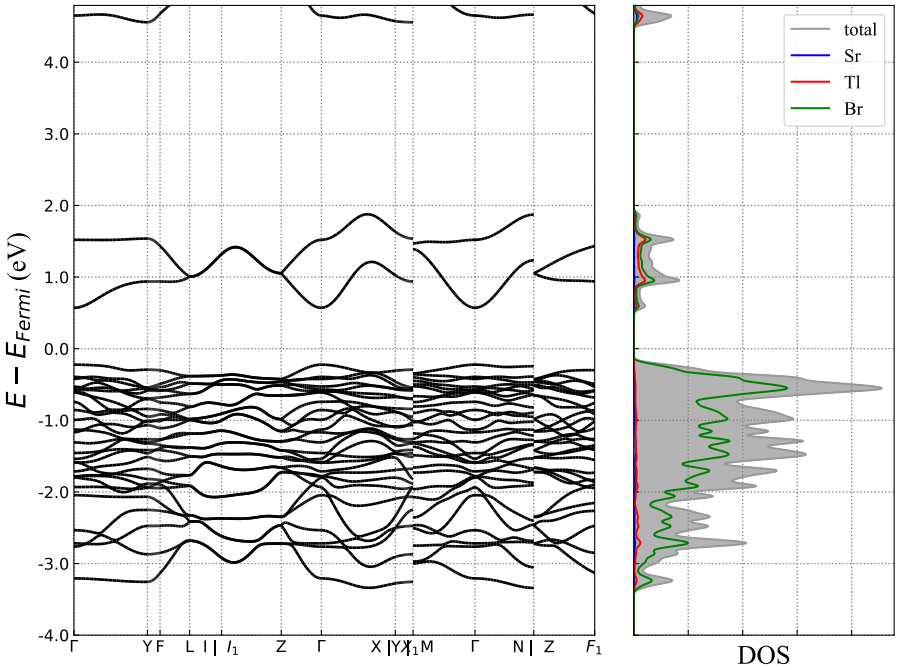
Fonte: elaborado pelo autor.

Figura B78: Estrutura de bandas e DOS do SrInF5.
Bandgap indireto de $E_g = 4.48$ eV.



Fonte: elaborado pelo autor.

Figura B79: Estrutura de bandas e DOS do SrTlBr5.
Bandgap direto de $E_g = 0.79$ eV.



Fonte: elaborado pelo autor.