UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA E ENGENHARIA DE MATERIAIS

INTEGRAÇÃO ALGORITMO GENÉTICO COM *MACHINE LEARNING* PARA DESIGN DE LIGAS DE ALTA ENTROPIA COM PROPRIEDADES MECÂNICAS OTIMIZADAS

Caroline Binde Stoco

São Carlos-SP
2025

UNIVERSIDADE FEDERAL DE SÃO CARLOS
CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA E ENGENHARIA DE MATERIAIS

INTEGRAÇÃO ALGORITMO GENÉTICO COM *MACHINE LEARNING* PARA DESIGN DE LIGAS DE ALTA ENTROPIA COM PROPRIEDADES MECÂNICAS OTIMIZADAS

Caroline Binde Stoco

Dissertação apresentada ao Programa de Pós-Graduação em Ciência e Engenharia de Materiais como requisito parcial à obtenção do título de MESTRA EM CIÊNCIA E ENGENHARIA DE MATERIAIS

Orientador: Dr. Francisco Gil Coury
Coorientador: Dr. Lucas Barcelos Otani
Agência Financiadora:
CAPES: Processo: 88887.843474/2023-00
FAPESP - Processos: 2023/04907-2 e 2023/14613-6

São Carlos-SP
2025

DEDICATÓRIA

A todos que acreditaram em mim, mesmo nos momentos em que eu mesma duvidei.

Obrigada.

VITAE

Bacharel em Engenharia de Materiais pela Universidade Federal de São Calos (2023).

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência e Engenharia de Materiais

---

Folha de Aprovação

Defesa de Dissertação de Mestrado da candidata Caroline Binde Stoco, realizada em 27/02/2025.

Comissão Julgadora:

Prof. Dr. Francisco Gil Coury (UFSCar)

Prof. Dr. Guilherme Zepon (UFSCar)

Prof. Dr. Witor Wolf (USP)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência e Engenharia de Materiais.

AGRADECIMENTOS

Primeiramente, eu gostaria de agradecer ao meu orientador, o professor doutor Francisco Gil Coury, ao meu co-orientador, o professor doutor Lucas Barcelos Otani e aos meus orientadores na França, os doutores Guillaume Deffrennes e Yannick Champion, por todo o auxílio durante a execução deste mestrado. Seus ensinamentos jamais serão esquecidos.

Ressalto também a importância dos técnicos de laboratório Edson Roberto D'Almeida e Rover Belo por me auxiliarem na parte experimental.

Agradeço também imensamente a todos que compõem o mezanino e seus agregados: Ana Soares, Anderson Fang, Argos Soares, Caio Gueiros, David Silva, Flávio Favaro, Gabriela Bugni, Guilherme Stumpf, Gustavo Bertoli, João Mota, Júlia Xaraba, Nicolas Moreira, Victor Hugo e Vinícius Bacurau. Vocês me auxiliaram tanto na parte intelectual, como experimental e o mais importante, no âmbito emocional. Os meus sinceros agradecimentos.

Esse trabalho jamais teria sido completo sem o apoio incondicional da minha família, especialmente a minha mãe Núbia de Cássia Binde Stoco e do meu namorado, João Vitor Franzin. Obrigada por estarem comigo nos momentos bons e não permitirem que eu ficasse presa nos momentos ruins.

Por fim, gostaria de agradecer as agências de fomento:

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

À CAPES/PROEX - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior/ Programa de Excelência Acadêmica pelo apoio financeiro para realização desse trabalho com bolsa de estudos, processo nº 88887.843474/2023-00.

À FAPESP - Fundação de Amparo à Pesquisa do Estado de São Paulo pelo apoio financeiro para realização desse trabalho com bolsa de estudos, processos 2023/04907-2 e 2023/14613-6.

RESUMO

O rápido avanço tecnológico tem impulsionado a necessidade de desenvolver novos materiais capazes de atender às demandas crescentes de diversos setores. Nesse contexto, as Ligas de Alta Entropia (HEAs) surgem como uma solução promissora. Entretanto, a seleção de composições ideais em um vasto e complexo espaço multicomposicional permanece um grande desafio. Para enfrentá-lo, este estudo desenvolveu um algoritmo genético capaz de realizar o design de HEAs otimizando múltiplos objetivos, potencialmente antagônicos. Por meio de processos de seleção, cruzamento e mutação genética, o algoritmo gerou novas gerações de ligas, alinhando progressivamente suas propriedades aos parâmetros desejados. O processo de otimização teve como objetivo obter uma estrutura monofásica cúbica de face centrada (CFC), avaliada pela integração do método CALPHAD com técnicas de *machine learning* para classificação utilizando *Support Vector Machine* (SVM) e *active learning*. Além disso, o algoritmo buscou maximizar a constante de Hall-Petch (K) e a tensão de cisalhamento crítica resolvida (τ_Y), aumentando a resistência mecânica por meio do refinamento de grão e do endurecimento por solução sólida. Esses parâmetros foram avaliados usando equações empíricas. Os efeitos da plasticidade induzida por maclação (TWIP) e plasticidade induzida por transformação de fase (TRIP) também foram incorporados ao algoritmo, com previsões da energia de falha de empilhamento (EFE) realizadas pelo modelo de regressão *Support Vector Regression* (SVR). O produto final deste algoritmo genético é um conjunto de composições otimizadas de HEAs, incluindo duas selecionadas para análise experimental futura, demonstrando a eficácia do algoritmo genético em explorar o espaço multicomposicional e lidar com objetivos de design conflitantes.

Palavras-chave: Ligas de Alta Entropia; Algoritmo Genético; Machine Learning; Método CALPHAD; Propriedades Mecânicas.

ABSTRACT

INTEGRATION OF GENETIC ALGORITHM WITH MACHINE LEARNING FOR THE DESIGN OF HIGH-ENTROPY ALLOYS WITH OPTIMIZED MECHANICAL PROPERTIES

The rapid pace of technological advancement is driving the need for the development of novel materials that can meet the evolving demands across various service sectors. In this context, High Entropy Alloys (HEAs) have emerged as a promising solution. However, a significant challenge has been the selection of optimal compositions within a vast and complex multicompositional space. To address this challenge, this study developed a genetic algorithm capable of designing HEAs by optimizing multiple, potentially antagonistic, objectives. Through processes of genetic selection, crossover and mutation, the algorithm generated new alloy generations, progressively aligning their properties with the desired parameters. The optimization process aimed to achieve a single-phase face-centered cubic (FCC) structure, assessed through the integration of the CALPHAD method with machine learning techniques for classification using Support Vector Machines (SVM) and active learning. Additionally, the algorithm sought to maximize the Hall-Petch constant (K) and the critical resolved shear stress (τ_Y), enhancing mechanical strength through grain refinement and solid solution strengthening. These parameters were evaluated using empirical equations. The effects of twinning-induced plasticity (TWIP) and transformation-induced plasticity (TRIP) were also incorporated into the algorithm, with stacking fault energy (SFE) predictions made using Support Vector Regression (SVR). The final output of this genetic algorithm is a set of optimized HEA compositions, including two selected for future experimental analysis, demonstrating the genetic algorithm effectiveness in navigating the multicompositional space and addressing conflicting design objectives.

Keywords: High Entropy Alloys; Genetic Algorithm; Machine Learning; CALPHAD Method; Mechanical Properties.

PUBLICAÇÕES

- C.B. Stoco, D.R. Cassar, G.L. Santana, M. Kaufman, A. Clarke, F.G. Coury, Optimizing Toughness in High Entropy Alloys Using a Genetic Algorithm: A Combined Computational and Experimental Approach, Mater. Today Commun. 41 (2024) 110768. https://doi.org/https://doi.org/10.1016/j.mtcomm.2024.110768.

ÍNDICE DE ASSUNTOS

Pág.

ÍNDICE DE TABELAS

ÍNDICE DE FIGURAS

SÍMBOLOS E ABREVIATURAS

Al	Alumínio
ANN	*Artificial Neural Networks*
ANNNI	*Axial Next-Nearest-Neighbor Ising*
$\boldsymbol{b}$	Magnitude do vetor de Burgers
CALPHAD	*Computer Coupling of Phase Diagrams and Thermochemistry*
CCC	Cúbica de Corpo Centrado
CFC	Cúbica de Face Centrada
Co	Cobalto
CCAs	*Complex Cencentrated Alloys*
Cr	Cromo
$\boldsymbol{D}$	Diâmetro de grão
$\boldsymbol{D_{gb}}$	Coeficiente de difusão do contorno de grão
$\boldsymbol{d_l}$	Tamanho de grão determinado pelo comprimento médio do intercepto linear
$\boldsymbol{d_s}$	Tamanho de grão espacial
DFT	*Density Functional Theory*
DFT-KKR-CPA	*DFT-Based Electronic-Structure Green's function method*
FAB	*Fully Assessed Binary systems*
FAT	*Fully Assessed Ternary systems*
Fe	Ferro
FSS	*Forward Sequential Selection*
GND	*Geometrically Necessary Dislocations*
HCP	Hexagonal Compacta
HEAs	*High Entropy Alloys*
IA	Inteligência Artificial
K	Constante de Hall-Petch
$\boldsymbol{k}$	Constante de Bolztman
KNN	K-vizinhos mais próximos
ML	*Machine Learning*

Mn	Manganês
MPEAs	*Multiprincipal Element Alloys*
Ni	Níquel
PBE	Perdew–Burke–Ernzerhof
poly	Polinomial
R	Constante universal dos gases
RBF	*Radial Basis Function*
RF	*Random Forest*
RHEA	*Refractory High Entropy Alloys*
RMSE	*Root Mean Squared Error*
SVM	*Support Vector Machine*
$\boldsymbol{T}$	Temperatura
TRIP	*Transformation Induced Plasticity*
TWIP	*Twinning Induced Plasticity*
V	Vanádio
$\boldsymbol{w_x}$	Peso da propriedade x na função *fitness*
$\boldsymbol{w_y}$	Peso da propriedade y na função *fitness*
x	Valor da propriedade x
$\boldsymbol{x_d}$	Valor desejado da propriedade x
y	Valor da propriedade y
$\boldsymbol{y_d}$	Valor desejado da propriedade y
$\boldsymbol{\alpha}$	Constante do modelo de Ashby
$\boldsymbol{\alpha'}$	Número adimensional
$\boldsymbol{\beta}$	Constante dependente do modelo de Hall-Petch
$\Delta \boldsymbol{E_b}$	Energia de ativação para mover uma discordância
$\Delta \boldsymbol{V_n}$	Volume médio de desajuste do soluto n
$\boldsymbol{\delta}$	Espessura do contorno de grão
$\boldsymbol{\varepsilon}$	Deformação plástica
$\dot{\boldsymbol{\varepsilon}}$	Taxa de deslizamento global
$\boldsymbol{\varepsilon_1},\ \boldsymbol{\varepsilon_2}, \boldsymbol{\varepsilon_3}$	Fatores de penalização
$\dot{\boldsymbol{\epsilon}}_{\boldsymbol{0}}$	Termo de referência para a taxa de deformação
$\dot{\boldsymbol{\epsilon}}$	Taxa de deformação aplicada ao material

$\boldsymbol{f_1(w_c)}$ e $\boldsymbol{f_2(w_c)}$	Campos de pressão adimensional
$\boldsymbol{\mu}$	Módulo de cisalhamento
$\boldsymbol{\rho}$	Densidade de discordâncias
$\boldsymbol{\sigma_0}$	Constante presente no modelo de Hall-Petch
$\boldsymbol{\sigma_y}$	Limite de escoamento
$\boldsymbol{\sigma_{\Delta V_n}}$	Desvio padrão devido a variações locais de composição
$\boldsymbol{\tau_0}$	Tensão de Peierls a 0K
$\boldsymbol{v}$	Coeficiente de Poisson

1 INTRODUÇÃO

As ligas de alta entropia (HEAs) ou ligas multicomponentes são uma nova classe de materiais metálicos caracterizados pela ausência de um único elemento principal, como ocorre em ligas convencionais de alumínio ou aço [1]. Essa característica permite maior flexibilidade no design de materiais, possibilitando o ajuste de propriedades específicas para diversas aplicações. No cenário atual, marcado pelo crescimento de setores, como o petroquímico, aeroespacial e naval, a demanda por materiais com propriedades otimizadas se torna cada vez mais relevante [2,3].

Devido ao vasto espaço composicional das HEAs, essas ligas podem ser projetadas para atingir uma ampla gama de propriedades, como alta resistência mecânica, resistência à corrosão, oxidação, fadiga e desgaste [2,3]. No entanto, a diversidade de composições possíveis dificulta o uso de abordagens tradicionais baseadas em tentativa e erro, elevando os métodos computacionais como ferramentas indispensáveis no design de ligas otimizadas [1]. Entre essas abordagens, destaca-se o uso de algoritmos genéticos, que permitem a otimização simultânea de múltiplos objetivos de forma eficiente [4].

Estudos anteriores conduzidos pela candidata resultaram no desenvolvimento de um algoritmo genético voltado para a seleção de ligas monofásicas com estrutura cúbica de face centrada (CFC) e propriedades mecânicas aprimoradas. Contudo, limitações foram identificadas na precisão da previsão das fases, já que o algoritmo dependia de parâmetros empíricos, como o parâmetro termodinâmico adimensional φ [5] e a concentração dos elétrons de valência (VEC) [6]. Testes experimentais indicaram que as composições selecionadas nem sempre resultavam em ligas monofásicas CFC, evidenciando a necessidade de melhorar a confiabilidade dos modelos preditivos.

Nesse contexto, propõe-se a substituição dos parâmetros empíricos por uma abordagem mais robusta, baseada na integração de algoritmos genéticos com cálculos CALPHAD de alto rendimento e estratégias de *machine learning*. O método CALPHAD é amplamente reconhecido por sua capacidade de prever diagramas de equilíbrio de fases em função de variáveis termodinâmicas [7]. Já o *machine learning*, como uma estratégia dentro da inteligência artificial, induz modelos preditivos com base em conjuntos de dados. Essa integração visa otimizar tanto a identificação de fases quanto a definição de propriedades mecânicas em ligas projetadas.

2 OBJETIVOS

Objetivo geral:

Desenvolver um algoritmo genético integrado ao *machine learning* e cálculos CALPHAD, capaz de otimizar a definição de composições para ligas de alta entropia, com foco em propriedades mecânicas avançadas e identificação de fases.

Objetivos específicos:

1. Obtenção de base de dados para previsão de fase;
2. Treinamento de modelo de *machine learning* para classificação de fases;
3. Treinamento de modelo de *machine learning* para previsão da energia de falha de empilhamento;
4. Adaptação do algoritmo genético com variáveis para previsão de fase e energia de falha de empilhamento e para cálculo da constante de Hall-Petch e da tensão de cisalhamento crítica resolvida;
5. Aprimoramento dos modelos de *machine learning* e do algoritmo genético;
6. Seleção de ligas de interesse.

3 REVISÃO DA LITERATURA

O primeiro tópico desta revisão bibliográfica se refere a uma descrição detalhada do material de interesse, ou seja, a respeito das ligas de alta entropia. Em seguida, serão abordados mecanismos de endurecimento presentes nessas ligas, em específico endurecimento por refino de grão e por solução sólida, e plasticidade induzida por maclagem e por transformação de fase. Por fim, serão descritos os métodos computacionais que serão empregados para a predição dos referidos mecanismos de endurecimento e das fases formadas, sendo esses o método CALPHAD, *machine learning* e algoritmo genético.

3.1 Ligas de Alta Entropia

Ligas de alta entropia (*High Entropy Alloys* – HEAs), também conhecidas como ligas de elemento multiprincipal (*Multiprincipal Element Alloys* – MPEAs), ligas multielemento ou ainda ligas complexas concentradas (*Complex Concentrated Alloys* – CCAs), representam uma classe inovadora de materiais metálicos. Essas ligas desafiam os princípios tradicionais de design de materiais ao combinar cinco ou mais elementos em proporções variando entre 5% e 35% em fração atômica. Diferentemente das ligas convencionais, cuja matriz geralmente é dominada por um elemento base, as HEAs derivam suas propriedades de uma combinação de múltiplos elementos principais [1,2,8,9].

Considerando que há 75 elementos estáveis na tabela periódica que não são tóxicos, radioativos ou gases nobres, calculam-se mais de 592 bilhões de composições de ligas considerando entre 3 a 6 elementos principais variando em concentrações de 10% [9]. Apesar de seu enorme potencial, a exploração do espaço composicional das HEAs continua sendo um desafio. A necessidade de desenvolver novas estratégias de experimentação e modelagem para navegar entre propriedades dependentes da composição e microestruturas projetadas é uma prioridade neste campo [1,9].

O desenvolvimento das HEAs busca explorar regiões centrais ainda pouco investigadas em diagramas de fases multicomponentes [3]. Primeiramente, focou-se em soluções sólidas monofásicas, como a liga de Cantor (CrMnFeCoNi), que apresenta estrutura cúbica de face centrada (CFC) e excelente resistência mecânica [10]. Outro exemplo notável é a liga equiatômica $Cr_{33}Co_{33}Ni_{33}$, uma das ligas mais tenazes já

desenvolvidas, combinando alta resistência e ductilidade, o que a torna promissora para aplicações estruturais [8,11]. Na Figura 1, observa-se que essas ligas ocupam a região superior direita de um mapa de Ashby, demonstrando uma combinação única de tenacidade à fratura e tensão de escoamento, equiparando-se às propriedades de aços criogênicos.

Além disso, a depender da composição, as HEAs apresentam notável resistência à oxidação, corrosão, fadiga e desgaste, mantendo valores aceitáveis de ductilidade. Essas propriedades excepcionais as tornam candidatas para aplicações em setores como aeroespacial, engenharia oceânica, nuclear, biomédico e indústria química [2,3].

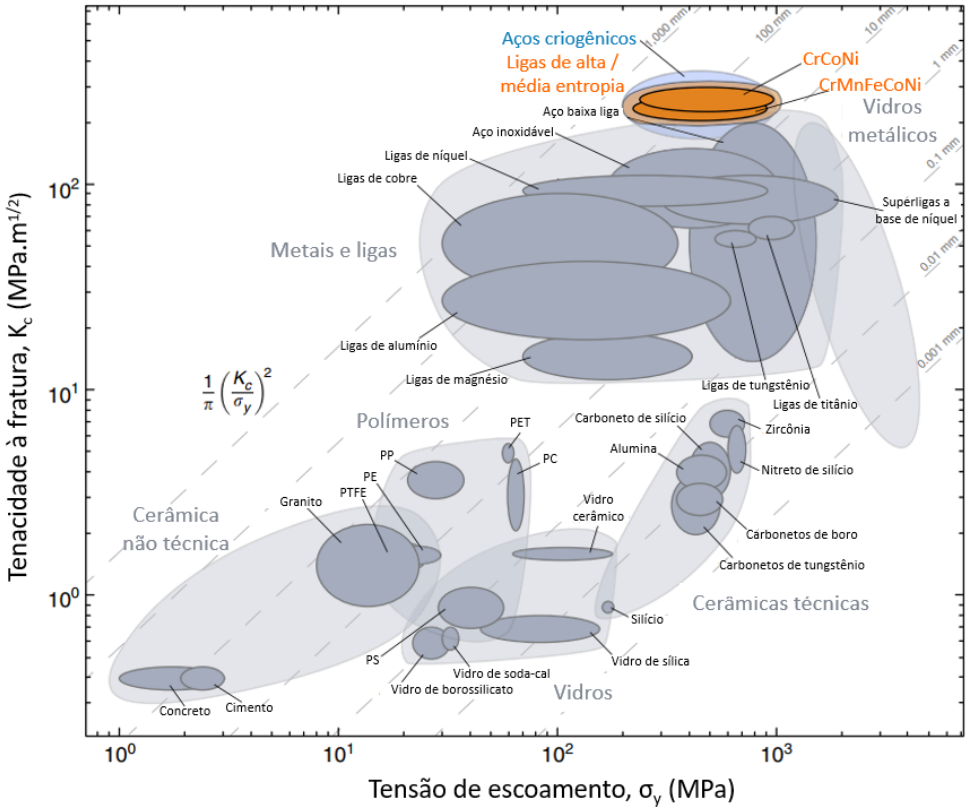

Figura 1 - Mapa de Ashby que correlaciona tenacidade à fratura e tensão de escoamento para diversos materiais [3] (adaptado).

As HEAs podem ser definidas com base em critérios de composição ou entropia configuracional. A definição composicional estabelece que as concentrações dos elementos principais variem entre 5% e 35%, sem impor limites à magnitude da entropia ou exigir a presença de soluções sólidas monofásicas. Já a definição baseada em entropia considera ligas de alta entropia aquelas cuja entropia configuracional de mistura excede 1,61 R, sendo R a constante universal dos gases [12,13].

A expressão “ligas de alta entropia” foi inicialmente adotada devido à ideia de maximizar a entropia configuracional de mistura em composições com um grande número de elementos em frações equiatômicas. Essa abordagem implica que a entropia configuracional de mistura pode atingir um máximo, favorecendo a estabilização de uma fase em solução sólida em detrimento de outras, como as intermetálicas [5]. Este foi o princípio utilizado para a idealização das já mencionadas liga de Cantor e da liga CrCoNi, por exemplo.

No entanto, essa concepção, que considera a entropia como indissociável da entalpia, foi posteriormente refutada [3,8]. Essa constatação levou ao surgimento de outras terminologias, como ligas de elemento multiprincipal, ligas multielemento ou ligas complexas concentradas. Contudo, o termo liga de alta entropia ainda persiste fortemente na literatura, sendo o mais usual e o adotado no decorrer do texto.

3.2 Mecanismos de endurecimento

3.2.1 Refino de grão

O refinamento de grãos é amplamente reconhecido como um dos mecanismos de endurecimento mais eficientes e desejáveis em materiais metálicos. Sua capacidade de aumentar significativamente os limites de escoamento e a resistência à tração, mantendo níveis aceitáveis de ductilidade, dentro de um regime de tamanho de grão onde a relação de Hall-Petch ainda é válida, o torna uma abordagem estratégica no design de ligas metálicas avançadas [8,14].

No início dos anos de 1950, Hall [15] e Petch [16] demonstraram empiricamente que o limite de escoamento, σ_y, de um material está relacionado com o seu tamanho de grão, D, de acordo com a Equação 1, sendo σ_0 e K constantes dependentes da composição química e microestrutura do material analisado. Inicialmente aplicado para aços, o modelo é atualmente amplamente utilizado para diversos materiais, desde metais puros, ligas diversas e até para estruturas mais complexas, como as cerâmicas [8].

$$\sigma_y = \sigma_0 + \frac{K}{\sqrt{D}} \quad (1)$$

Ao longo dos anos, diversos modelos foram propostos visando correlacionar aspectos físicos para os termos presentes na Equação 1. O modelo clássico do empilhamento de discordâncias considera que os contornos de grão atuam como barreiras

ao movimento das discordâncias, que se acumulam ao longo dos planos de escorregamento até que a soma da tensão externa aplicada e da tensão gerada na ponta do empilhamento seja suficiente para superar essa barreira e permitir o deslizamento para o próximo grão. Esse mecanismo implica que o comprimento do empilhamento de discordâncias é proporcional ao tamanho de grão. Portanto, grãos menores resultam em empilhamentos mais curtos, o que reduz a tensão na extremidade do empilhamento e aumenta a tensão externa necessária para ativar o deslizamento no grão adjacente. Consequentemente, a relação entre a tensão de escoamento e o inverso da raiz quadrada do tamanho de grão segue a forma linear descrita pela equação de Hall-Petch [17,18].

Há diversas variações desse modelo, conforme revisado por Li e Chou [19], mas todos mantêm a mesma forma básica apresentada na Equação 2.

$$\sigma_y = \sigma_0 + \beta\mu\sqrt{\frac{b}{D}} \tag{2}$$

Nessa equação, β é uma constante dependente do modelo (em média 0,18 para materiais cúbicos de face centrada - CFC, 0,42 para materiais cúbicos de corpo centrado - CCC e 0,9 para materiais hexagonais compactos - HCP), μ é o módulo de cisalhamento, b é a magnitude do vetor de Burgers de discordâncias móveis e D o tamanho do grão [17]. Nota-se que, pela Equação 2, a constante de Hall-Petch K pode ser descrita conforme Equação 3. Ela deve ser entendida como a resistência oferecida pelos contornos de grão ao movimento das discordâncias. Um valor elevado de K indica que a resistência ao escoamento aumenta significativamente à medida que o tamanho dos grãos diminui, evidenciando a eficácia do refino de grão como mecanismo de fortalecimento [14].

$$K = \beta\mu\sqrt{b} \tag{3}$$

Um segundo modelo foi proposto por Ashby (1970) [20], baseado no conceito de discordâncias geometricamente necessárias (*Geometrically Necessary Dislocations* - GNDs), e destaca a dependência do coeficiente de Hall-Petch K com a deformação plástica. Esse modelo foi construído a partir de duas observações principais:

1. A densidade de discordâncias ρ cresce linearmente com a deformação plástica e com o inverso do tamanho de grão durante a deformação uniforme de materiais policristalinos [17].
2. Mesmo quando uma amostra policristalina é submetida a uma deformação plástica uniforme macroscópica, os grãos individuais apresentam um fluxo plástico não

homogêneo. Isso ocorre devido às restrições de compatibilidade nos contornos de grão, já que a deformação plástica em um único sistema de escorregamento só promove mudanças geométricas específicas. Caso apenas um sistema de escorregamento esteja ativo, problemas de compatibilidade surgem nos contornos de grãos, exigindo a ativação de sistemas de escorregamento adicionais. Isso gera um excesso de discordâncias próximas aos contornos de grão para restaurar a compatibilidade [17]. Esse processo é ilustrado na Figura 2.

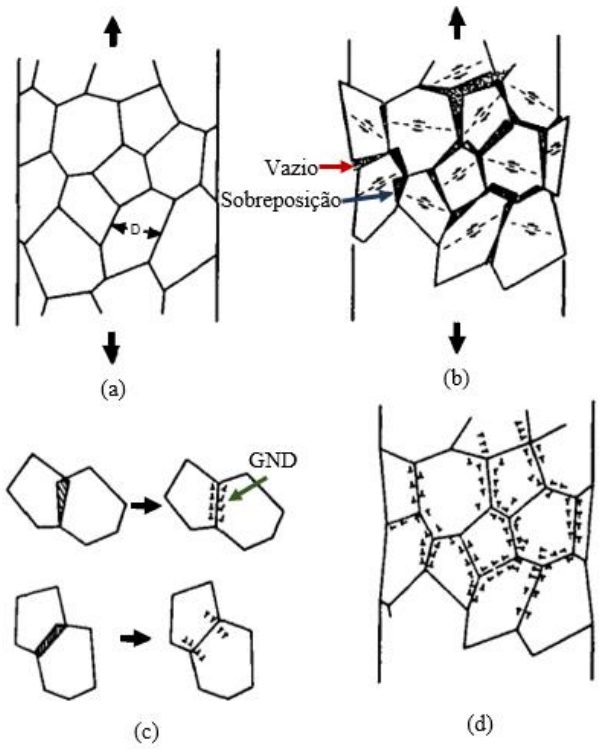

Figura 2 - Esquema de formação de GND. (a) Início da deformação de um corpo policristalino; (b) Formação de vazios e sobreposições de grãos devido à deformação; (c) e (d) Criação de GND para compatibilização dos contornos de grão e eliminação de defeitos [20] (adaptado).

Para restaurar a compatibilidade nos contornos de grão, Ashby estimou que a densidade das GND devem ser proporcionais a $\varepsilon/4Db$, sendo ε a deformação plástica macroscópica, D o diâmetro de grão e b a magnitude do vetor de Burgers [17]. Dessa forma, o modelo de Ashby (Equação 4) explica o modelo de Hall-Petch ao combinar suas predições sobre a dependência da deformação e do tamanho de grão com a densidade de discordâncias, junto com o modelo de endurecimento de Taylor. Nesse modelo, o termo α é uma constante e assume-se que a densidade de GNDs é consideravelmente superior à densidade das demais discordâncias presentes no material [17].

$$\sigma_y \sim \sigma_0 + \alpha\mu\sqrt{\frac{b\varepsilon}{4D}} \quad (4)$$

Estudos subsequentes confirmaram que esse modelo é o mais consistente com observações experimentais de materiais policristalinos de grãos grosseiros [17,18]. Assim, o modelo de Ashby não apenas reforça a relação entre o tamanho de grão e a resistência mecânica, mas também evidencia que o coeficiente de Hall-Petch pode variar com a deformação plástica, destacando a relevância das GNDs no endurecimento de materiais.

Atualmente, Figueiredo et al. (2023) [18] propôs uma relação que unifica o efeito do tamanho de grão em baixas e altas temperaturas, baseada no modelo de deslizamento de contornos de grão. Esse modelo considera que as discordâncias extrínsecas — defeitos lineares de não equilíbrio gerados durante a deformação plástica pela interação de discordâncias da rede cristalina com os contornos de grão [20,21] — deslizam nos contornos de grão, promovendo o deslizamento desses contornos e gerando tensões em junções triplas. Essas tensões elevadas ativam novos deslizamentos nos grãos vizinhos, levando à emissão de discordâncias dos contornos de grão, que deslizam através do grão e se acumulam no contorno oposto. Posteriormente, essas discordâncias são absorvidas pelo contorno oposto por meio de escalagem, um processo auxiliado pelos altos níveis de tensão desenvolvidos no empilhamento de discordâncias. Esse fenômeno é ilustrado na Figura 3. Reduzir o tamanho de grão diminui o comprimento do empilhamento e aumenta a tensão necessária para o fluxo plástico, consistente com a relação de Hall-Petch.

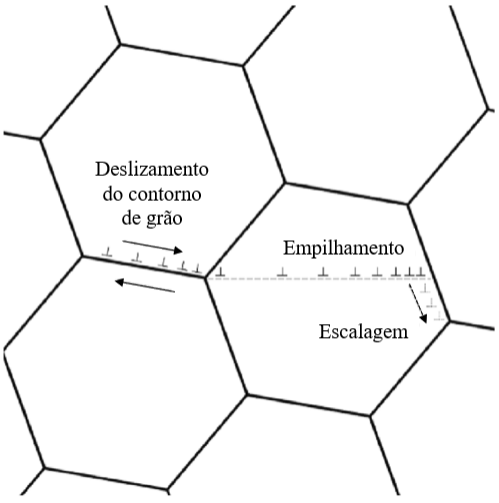

Figura 3 - Esquema dos princípios do mecanismo de deslizamento do contorno de grão [18] (adaptado).

Figueiredo et al. (2023) [18] desenvolveram as Equações 5 e 6 para descrever a relação entre a tensão de escoamento σ e o tamanho de grão, considerando diversos parâmetros para altas e baixas temperaturas, respectivamente. Nas equações, k é a constante de Bolztman, T é a temperatura, d_s é o tamanho de grão espacial, d_l é o tamanho de grão determinado pelo comprimento médio do intercepto linear, $\dot{\varepsilon}$ é a taxa de deslizamento global, δ é a espessura do contorno de grão e D_{gb} é o coeficiente de difusão do contorno de grão.

$$\sigma \sim \sqrt{\frac{3\mu kT}{2d_s b^2} \ln\left(\frac{\dot{\varepsilon} {d_s}^3}{10\delta D_{gb}} + 1\right)} \tag{5}$$

$$\sigma \sim \sqrt{\frac{3\mu kT}{2d_l b^2} \ln\left(\frac{\dot{\varepsilon} {d_l}^3}{2\delta D_{gb}} + 1\right)} \tag{6}$$

A seguir (Tabela 1), é possível conferir um resumo das equações apresentadas nesta seção.

Tabela 1 - Modelos que visam explicar a relação observada por Hall-Petch

Modelo	Equação	Referência
Empilhamento de discordâncias	$\sigma_y = \sigma_0 + \beta\mu\sqrt{\frac{b}{D}}$	[17]
Discordâncias geometricamente necessárias	$\sigma_y \sim \sigma_0 + \alpha\mu\sqrt{\frac{b\varepsilon}{4D}}$	[20]
Deslizamento de contornos de grão	Altas temperaturas: $\sigma \sim \sqrt{\frac{3\mu kT}{2d_s b^2} \ln\left(\frac{\dot{\varepsilon} {d_s}^3}{10\delta D_{gb}} + 1\right)}$ Baixas temperaturas: $\sigma \sim \sqrt{\frac{3\mu kT}{2d_l b^2} \ln\left(\frac{\dot{\varepsilon} {d_l}^3}{2\delta D_{gb}} + 1\right)}$	[18]

3.2.2 Solução sólida

Uma solução sólida é formada quando, ao adicionar átomos de soluto ao material hospedeiro, a estrutura cristalina é mantida, sem a formação de novas fases cristalinas [22]. Essas soluções podem ser classificadas como substitucionais, quando os átomos de

soluto substituem predominantemente os do solvente na rede cristalina, ou intersticiais, quando ocupam majoritariamente posições nos interstícios dessa rede [23].

Os fatores que favorecem a formação de uma solução sólida substitucional foram amplamente estudados em trabalhos clássicos de Hume-Rothery [24]. Para uma maior solubilidade, recomenda-se que a diferença de tamanho atômico entre os átomos seja inferior a 15%, que ambos os elementos possuam a mesma estrutura cristalina em estado puro, a mesma valência e valores de eletronegatividade similares.

A introdução de átomos de soluto e a formação de uma solução sólida geralmente resultam no aumento da dureza e da tensão de escoamento do material. Essa melhoria nas propriedades mecânicas está relacionada à interação elástica entre os campos de deformação ao redor dos átomos de soluto e as discordâncias na rede cristalina. O endurecimento é proporcional à distorção da rede causada pela presença do soluto [23].

Além disso, outros mecanismos de interação entre soluto e discordâncias incluem:

- Interação de módulo: alterações locais no módulo de cisalhamento μ devido à presença do soluto;
- Interação por falha de empilhamento: segregação preferencial dos átomos de soluto nas falhas de empilhamento delimitadas por discordâncias parciais;
- Interações elétricas: átomos de soluto com cargas localizadas interagem com discordâncias que apresentam dipolos elétricos;
- Interações de curta ordem: atração preferencial entre átomos de soluto e seus vizinhos mais próximos;
- Interações de longa ordem: em ligas que formam superredes, o movimento de discordâncias pode gerar contornos de anti-fase [19].

Varvenne et al. [25,26] desenvolveram uma teoria mecanicista para o limite de escoamento em ligas de estrutura cúbica de face centrada com composições arbitrárias. Nesse modelo, cada elemento da liga é tratado como um soluto em uma matriz com propriedades médias. O endurecimento observado em ligas em comparação a metais puros é atribuído à formação de uma solução sólida com 100% de concentração de soluto. Nesse contexto, o aumento da energia necessária para o movimento de uma discordância reflete a soma das energias de interação entre os átomos de soluto e uma discordância individual. As principais contribuições para essa energia de interação vêm:

- Da interação elástica entre o campo de tensão da discordância e o tensor de deformação gerado pela distorção da rede cristalina causada pelo soluto.
- De interações químicas, que ocorrem quando os átomos de soluto estão localizados nos núcleos parciais das discordâncias ou ao longo das falhas de empilhamento.

Para descrever o limite de escoamento em diferentes condições, as Equações 7 e 8 foram desenvolvidas por Varvenne et al. [25,26] para baixas temperaturas e/ou altas tensões e altas temperaturas e/ou baixas tensões, respectivamente. Nessas equações, τ_0 é a tensão de Peierls a 0K; k é a constante de Boltzmann; T é a temperatura; ΔE_b é a energia de ativação para mover uma discordância; $\dot{\epsilon_0}$ é um termo de referência para a taxa de deformação, sendo proposto como $10^{-4}s^{-1}$ por Varvenne et al. [25,26] e $\dot{\epsilon}$ é a taxa de deformação aplicada ao material, usualmente igual a $10^{-3}s^{-1}$ em ensaios de tração de materiais metálicos [27].

$$\tau_Y(T,\dot{\epsilon}) = \tau_0 \left[1 - \left(\frac{kT}{\Delta E_b} ln \frac{\dot{\epsilon_0}}{\dot{\epsilon}}\right)^{\frac{2}{3}}\right] \tag{7}$$

$$\tau_Y(T,\dot{\epsilon}) = \tau_0 \exp\left(\frac{-1}{0.51}\frac{kT}{\Delta E_b} ln \frac{\dot{\epsilon_0}}{\dot{\epsilon}}\right) \tag{8}$$

τ_0 e ΔE_b, por sua vez, podem ser calculados de acordo com as Equações 9 e 10 respectivamente, sendo α' um número adimensional igual a 0,123, μ o módulo de cisalhamento, b a magnitude do vetor de Burgers, v o coeficiente de Poisson, $f_1(w_c)$ e $f_2(w_c)$ campos de pressão adimensional igual a 0,35 e 5,70, respectivamente, ΔV_n o volume médio de desajuste do soluto n (variação do volume de uma célula unitária CFC composta por um único elemento em comparação com o seu volume médio considerando a composição da liga) e $\sigma_{\Delta V_n}$ o desvio padrão devido a variações locais de composição (termo de difícil obtenção e que pode ser desconsiderado) [25,26].

$$\tau_0 = 0.051\alpha'^{-\frac{1}{3}}\mu\left(\frac{1+v}{1-v}\right)^{\frac{4}{3}} f_1(w_c) \times \left[\frac{\sum_n c_n\left(\Delta\bar{V}_n{}^2 + \sigma_{\Delta V_n}{}^2\right)}{b^6}\right]^{\frac{2}{3}} \tag{9}$$

$$\Delta E_b = 0.274\alpha'^{\frac{1}{3}}\mu b^3\left(\frac{1+v}{1-v}\right)^{\frac{2}{3}} f_2(w_c) \times \left[\frac{\sum_n c_n\left(\Delta\bar{V}_n{}^2 + \sigma_{\Delta V_n}{}^2\right)}{b^6}\right]^{\frac{1}{3}} \tag{10}$$

Dessa forma, a teoria postula que a resistência do material, refletida em seu limite de escoamento, não está diretamente relacionada ao número de componentes, nem é maximizada em composições equiatômicas necessariamente. Em vez disso, ela pode ser otimizada ao maximizar o desajuste volumétrico médio quadrático ponderado pela concentração e/ou aumentar o módulo de cisalhamento médio do sistema [25,26].

3.2.3 Plasticidade induzida por maclagem e por transformação de fase

3.2.3.1 Energia de falha de empilhamento

As estruturas cristalinas cúbicas de face centrada (CFC) e hexagonais compactas (HCP) podem ser descritas pelo empilhamento de planos atômicos compactos. A principal diferença entre essas estruturas está na sequência de empilhamento: na estrutura CFC, cujos planos compactos pertencem à família (111), o empilhamento segue o padrão ABCABCABC... (Figura 4.a), enquanto na estrutura HCP, cujos planos compactos pertencem à família (0001), o padrão é ABABAB... (Figura 4.b) [22].

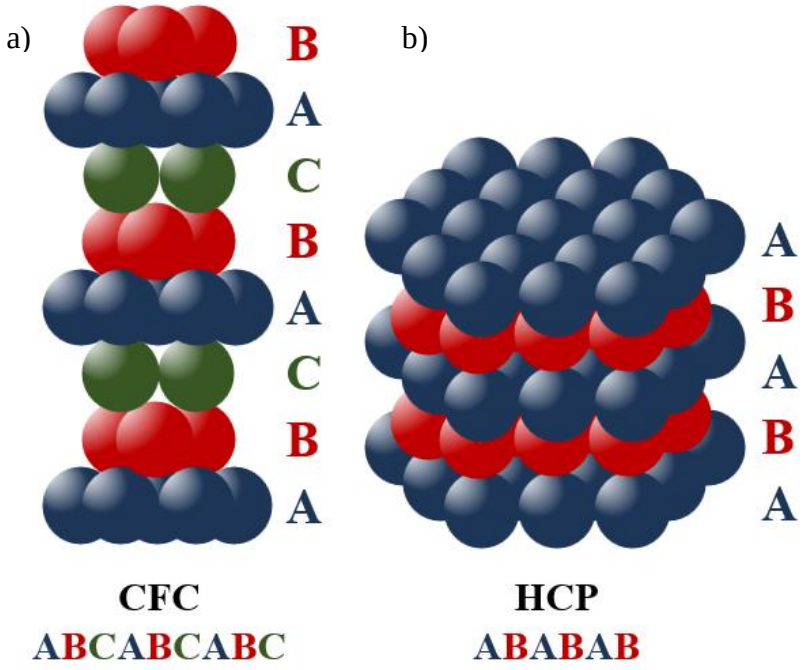

Figura 4 - Sequência de empilhamento de planos atômicos compactos para a estrutura cristalina (a) cúbica de face centrada e (b) hexagonal compacta.

Uma falha de empilhamento é um defeito planar que interrompe a ordem regular do empilhamento dos planos atômicos, causando uma alteração local na estrutura

cristalina. Em cristais CFC, podem ocorrer dois tipos de falhas de empilhamento: intrínsecas e extrínsecas. Falhas intrínsecas resultam da remoção de uma camada da sequência de empilhamento, mantendo a ordem acima e abaixo do defeito (Figura 5.a). Já nas falhas extrínsecas, ocorre a adição de uma camada extra, quebrando a sequência de empilhamento em dois pontos, acima e abaixo do defeito (Figura 5.b). Em ambas as falhas, a sequência interna do empilhamento (que em um reticulado perfeito seria ABC) torna-se localmente AB, caracterizando uma estrutura HCP [28].

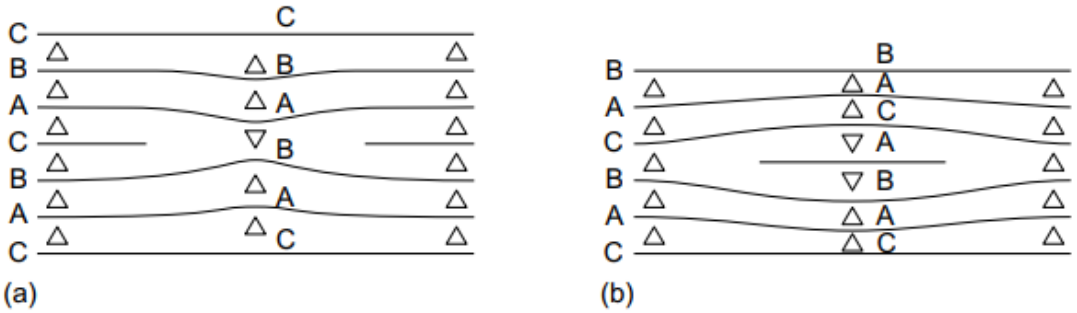

Figura 5 - Falhas de empilhamento (a) intrínseca e (b) extrínseca em uma estrutura cúbica de face centrada. Planos em relação normal entre si são separados por Δ, enquanto planos com erro de empilhamento são separados por ▽ [28].

Falhas de empilhamento em cristais CFC podem ser formadas através da dissociação de discordâncias perfeitas $\frac{1}{2}[110]\{111\}$ em discordâncias parciais de Shockley $\frac{1}{6}[112]\{111\}$ ou por meio da formação de discordâncias parciais de Frank [29]. O movimento das discordâncias parciais de Shockley, em modos específicos, são os principais responsáveis pela ocorrência dos efeitos TWIP e TRIP explicados adiante.

Como os átomos localizados em ambos os lados de uma falha de empilhamento não ocupam as posições esperadas em um reticulado perfeito, uma energia de superfície é gerada. Essa energia, combinada com outros fatores, como a afinidade do material em adotar uma estrutura HCP, definem a chamada energia de falha de empilhamento (EFE) [30]. A energia de falha de empilhamento proporciona uma força γ por unidade de comprimento de linha de discordância, tendendo a atrair as discordâncias parciais. A separação entre essas discordâncias é determinada pelo equilíbrio entre a força repulsiva das discordâncias e a energia de superfície associada à falha [23,28]. Valores baixos de EFE permitem maior separação entre as discordâncias parciais, enquanto valores elevados resultam em falhas mais próximas ou até inibem a sua formação [28].

Em materiais CFC com baixa EFE, a maior separação entre as discordâncias parciais de Shockley dificulta o deslizamento cruzado dessas discordâncias. Quando esse mecanismo de deformação é dificultado, outros mecanismos tornam-se predominantes, como o TWIP e o TRIP, beneficiando a ductilidade do material. [28].

3.2.3.2 Efeitos TWIP e TRIP

Os mecanismos de deformação plástica em materiais dependem diretamente da composição química e das condições de temperatura de deformação. Entre esses mecanismos, destacam-se o deslizamento de discordâncias, assim como a maclação e a transformação de fase adifusional. A redução da energia de falha de empilhamento promove uma transição nos mecanismos de deformação: inicialmente, o deslizamento de discordâncias predomina; em seguida, ocorre o deslizamento combinado com maclação; e, por fim, deslizamento combinado com transformação de fase martensítica [31].

A transformação martensítica induzida por deformação ocorre tipicamente em materiais com valores baixos de EFE, inferiores a 20 mJ/m². Por outro lado, a maclação induzida por deformação é observada em materiais com valores intermediários de EFE, entre 20-40 mJ/m². Valores de EFE superiores a 45 mJ/m² favorecem predominantemente a deformação plástica por deslizamento de discordâncias [31].

O efeito TWIP (*Twinning Induced Plasticity*), ou plasticidade induzida por maclação, consiste na formação de maclas de deformação que atuam como mecanismos de acomodação de tensão. Essa formação ocorre pela passagem de uma parcial de Shockley $b = \frac{1}{6}\langle 112\rangle$ em três planos sucessivos {111} (Figura 6.a) [31,32]. As maclas criam barreiras adicionais ao movimento de discordâncias, aumentando significativamente a capacidade de encruamento e a ductilidade do material [23].

Por sua vez, o efeito TRIP (*Transformation Induced Plasticity*), ou plasticidade induzida por transformação de fase, está associado à transformação martensítica durante a deformação. Em ligas multicomponentes, o efeito TRIP frequentemente se manifesta como a transformação da estrutura cristalina cúbica de face centrada (CFC) para hexagonal compacta (HCP). Essa transformação resulta da passagem de uma parcial de Shockley $b = \frac{1}{6}\langle 112\rangle$ em planos {111} alternados (Figura 6.b) [31,32]. Materiais que apresentam o efeito TRIP podem estar associados a uma excelente combinação de

resistência, ductilidade, dureza e propriedades em fadiga [29]. A ativação tanto do efeito TWIP como TRIP é fortemente influenciada pela composição química e pela estabilidade termodinâmica das fases envolvidas [28].

a)

b)

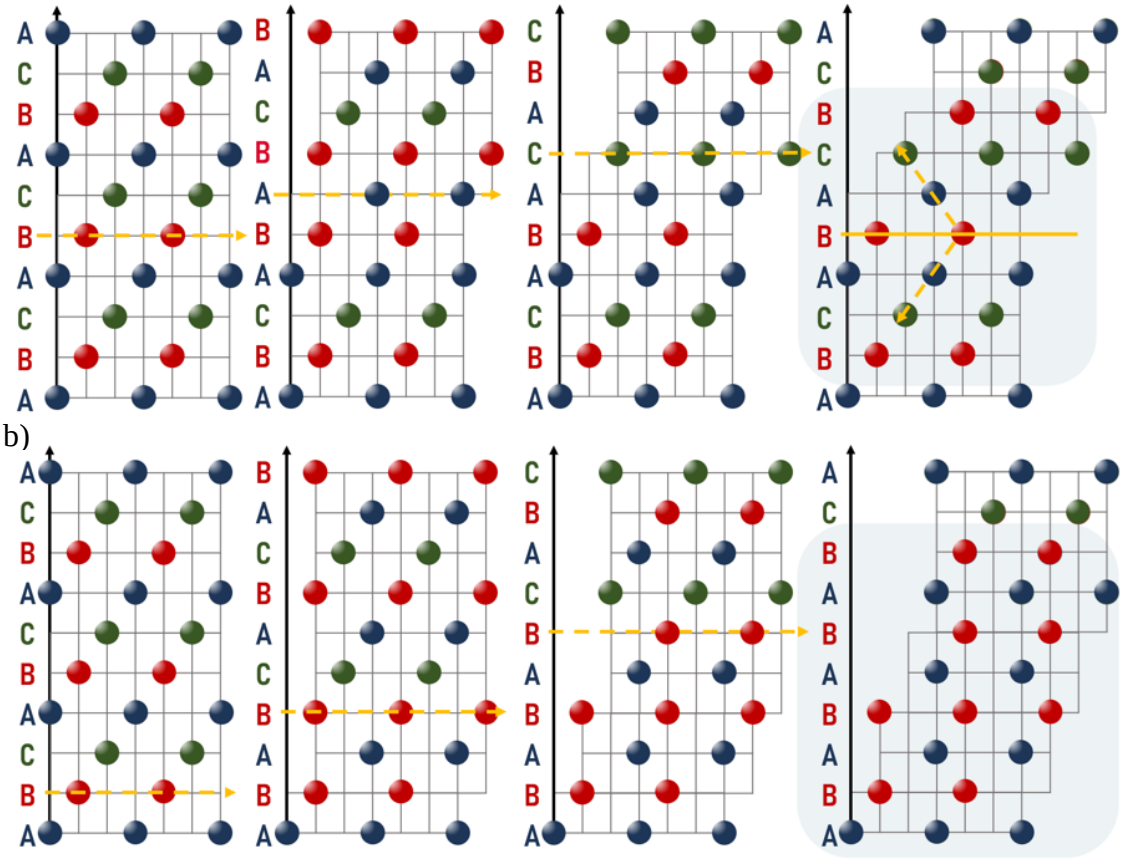

Figura 6 - Esquemática para formação de (a) maclas e (b) estrutura HCP em um cristal CFC por meio do deslocamento de parciais de Shockley.

Em determinadas condições, a formação extensiva de nano-maclas e lamelas HCP pode desencadear um efeito dinâmico de Hall-Petch, devido à criação de interfaces adicionais [14]. Essas interfaces atuam como barreiras ao movimento de discordâncias, reduzindo seu caminho livre médio e aumentando a taxa de encruamento. Segundo o critério de Considère, esse comportamento retarda o início da instabilidade plástica por empescoçamento, resultando em materiais com uma combinação excepcional de resistência e ductilidade, essenciais para aplicações estruturais críticas [14,29].

Dessa forma, ressalta-se a importância do design de ligas de alta entropia cujas composições resultem em materiais de baixa EFE, com tendência à ativação dos mecanismos TWIP e/ou TRIP.

3.2.3.2 Cálculo da energia de falha de empilhamento pela teoria do funcional da densidade

Uma das formas de se calcular a energia de falha de empilhamento é por meio da teoria do funcional da densidade (*Density Functional Theory* – DFT). A seguir, será apresentado o estudo de Khan et al. [33], no qual os autores calcularam a EFE de 489 ligas compostas pelos elementos Al, Co, Cr, Fe, Mn, Ni e V por meio de DFT, com o objetivo de identificar composições adequadas para aplicações em alta temperatura e manufatura aditiva.

Neste estudo, inicialmente, 1.000.000 de amostras foram geradas no espaço composicional CoCrFeMnNiVAl, das quais 467.228 composições com até 10% atômico de Al foram selecionadas. Após cálculos CALPHAD, 36.294 ligas monofásicas CFC a 1.073 K foram identificadas, considerando-se a estabilidade até a fusão. Critérios adicionais (temperatura *solidus* acima de 1.600 K e intervalo de solidificação menor que 100 K) reduziram o conjunto a 20.147 ligas, refinadas para 5.379 composições adequadas a altas temperaturas via simulações Scheil. Para enriquecer a base, 398 ligas foram selecionadas, acrescidas de 100 adicionais para maior cobertura composicional [33].

Os valores de energia de falha de empilhamento para as composições finais foram então calculados por meio da teoria do funcional da densidade. Foi utilizado o modelo axial de Ising do vizinho mais próximo (*Axial Next-Nearest-Neighbor Ising* – ANNNI) para descrever a sequência de empilhamento. A energia total do modelo ANNNI foi obtida pelo método da função de Green com estrutura eletrônica baseada em DFT (DFT-*Based Electronic-Structure Green's function method* – DFT-KKR-CPA). Para os cálculos, foi empregado o funcional de Perdew–Burke–Ernzerhof (PBE) e o método de pontos-k de Monkhorst–Pack para a integração da zona de Brillouin. Além disso, devido à presença de elementos magnéticos como Fe, Co, Cr, Ni e Mn, foi aplicada a configuração de polarização de spin

Ressalta-se que trabalhar diretamente com DFT não faz parte do escopo deste trabalho; apenas as informações principais da metodologia de cálculo realizada por terceiros foram incluídas. Para detalhes adicionais e maior entendimento como um todo da técnica de DFT, os leitores podem consultar o trabalho de Khan et al. [33] e outras referências específicas como [34,35].

A distribuição dos elementos na base de dados criada por [33] está representada na Figura 7 por meio de um gráfico violino. Observa-se que, com exceção de Al e V, cuja faixa de concentração na base de dados está limitada a valores abaixo de 15% atômico, os demais elementos apresentam distribuições que podem alcançar teores superiores a 35% atômico.

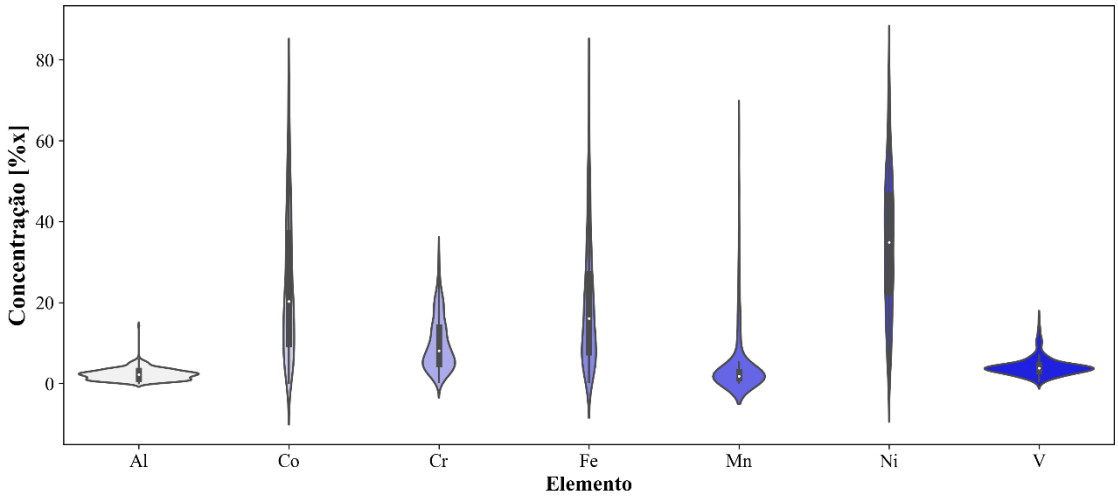

Figura 7 - Gráfico violino da distribuição de concentração de cada um dos elementos em porcentagem atômica na base de dados de energia de falha de empilhamento.

3.3 Método CALPHAD

O avanço no desenvolvimento de ligas metálicas exige um conhecimento aprofundado sobre suas propriedades, especialmente no que diz respeito às fases em equilíbrio em função de diversas variáveis de estado [7,36].

Essas informações são frequentemente obtidas por meio de diagramas de fases, que indicam as fases em equilíbrio em função da composição e da temperatura (T), sob pressão (P) constante. O estado de equilíbrio sob uma certa T e P é alcançado quando a energia livre de Gibbs do sistema atinge seu valor mínimo [36].

Historicamente, a construção desses diagramas envolvia extensos estudos experimentais, demandando significativos recursos de tempo e dinheiro, o que dificultava sua ampla aplicação [7,36]. O método CALPHAD (*Computer Coupling of Phase Diagrams and Thermochemistry*) surgiu como uma alternativa para superar esses desafios. Em essência, esse método baseia-se na determinação da energia livre de Gibbs para todas as possíveis fases presentes em um determinado sistema. Essas funções

dependem de variáveis de estado (como pressão, temperatura e composição) e de parâmetros empíricos específicos para cada fase. Uma vez determinadas, essas funções são reunidas em bases de dados termodinâmicas que podem ser utilizadas para calcular o equilíbrio de fases, propriedades termodinâmicas e diagramas de sistemas complexos. Na prática, após a avaliação dos parâmetros de interação em sistemas binários e ternários, torna-se viável estimar sistemas de ordens superiores por meio de extrapolação, sem requerer a inclusão de parâmetros empíricos adicionais [7].

Apesar de suas vantagens, o método CALPHAD enfrenta desafios significativos quando aplicado a HEAs, principalmente devido à dependência da qualidade das descrições e extrapolações baseadas em sistemas binários e ternários [37]. Uma das limitações mais críticas do método é a sua incapacidade de prever a estabilidade de fases em sistemas multicomponentes que não aparecem em sistemas de ordem inferior. Essa restrição tem sido parcialmente superada por meio da integração de cálculos termodinâmicos com abordagens *ab initio*, como o método DFT, que permite prever a estabilidade de fases concorrentes e enriquecer os cálculos de equilíbrio [7,38,39].

Dessa forma, para garantir a credibilidade dos cálculos realizados pelo método CALPHAD, é essencial considerar a qualidade dos dados termodinâmicos disponíveis, principalmente no caso de sistemas multicomponentes. Um dos critérios mais importantes para avaliar a confiabilidade dos cálculos CALPHAD é a fração de sistemas binários e ternários que foram completamente avaliados termodinamicamente, conhecidos como FAB (*Fully Assessed Binary systems*) e FAT (*Fully Assessed Ternary systems*). Esses critérios são fundamentais para garantir que as bases de dados utilizadas nos cálculos sejam robustas e reflitam as interações e fases do sistema [40–42].

3.4 *Machine Learning*

A Inteligência Artificial (IA) abrange técnicas que permitem a simulação de comportamentos humanos por computadores, visando superar a capacidade humana em decisões complexas de maneira independente ou com mínima intervenção. Entre os problemas centrais abordados pela IA, destacam-se representação de conhecimento, raciocínio, aprendizado, planejamento, percepção e comunicação [43].

O *machine learning* (ML), uma subárea da Inteligência Artificial, tem se consolidado como uma ferramenta poderosa na análise de dados e na automação de

processos decisórios. Ao invés de depender de regras explícitas de programação, o ML permite que sistemas "aprendam" com os dados, identificando padrões e criando modelos preditivos com base em exemplos fornecidos durante o treinamento [43].

O aprendizado pode ser categorizado em três tipos principais: supervisionado, não supervisionado e por reforço. No aprendizado supervisionado, o modelo é treinado com um conjunto de dados que contém tanto as entradas quanto as saídas desejadas (rótulos). Este tipo de aprendizado é amplamente utilizado em problemas de classificação e regressão. No aprendizado não supervisionado, os dados não são rotulados, e o modelo tenta identificar padrões ou agrupamentos. O aprendizado por reforço, por outro lado, envolve a interação com um ambiente onde o modelo aprende a tomar decisões por meio de tentativa e erro, recebendo recompensas ou punições com base nas ações que executa, sendo útil para problemas de otimização contínua [43–45].

O processo de treinamento de um modelo de ML pode ser dividido em várias etapas cruciais: Seleção de dados, engenharia de características, seleção do modelo, treinamento e validação [44]. Uma esquemática do processo pode ser conferida na Figura 8.

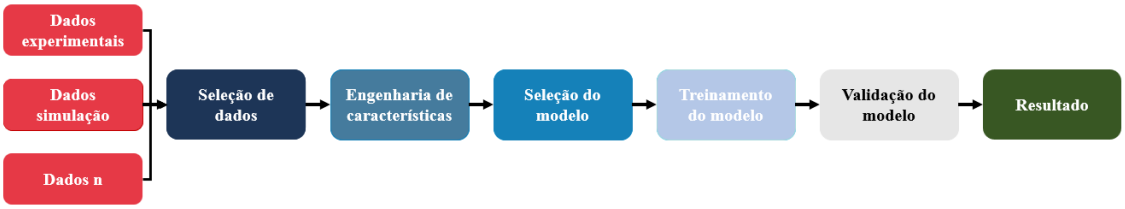

Figura 8 - Fluxograma simplificado da técnica de *Machine Learning.*

Seleção de dados: A seleção de dados envolve a escolha criteriosa dos atributos com base em seu tipo, qualidade e formato, priorizando sempre fontes confiáveis. Dados de alta qualidade são essenciais para evitar a inclusão de informações redundantes, incompletas ou errôneas, que poderiam comprometer os resultados das análises. Na área de ciência dos materiais, esses dados podem ser obtidos tanto por experimentos quanto por simulações e frequentemente abrangem informações físicas, químicas, estruturais e termodinâmicas dos materiais [43,44,46].

Engenharia de características: A engenharia de características consiste em derivar propriedades relevantes a partir dos dados brutos, com o objetivo de representá-los de forma adequada para a construção de modelos preditivos. Este processo busca preservar

informações discriminantes essenciais para a tarefa de aprendizado, permitindo que os modelos explorem os aspectos mais significativos dos dados para realizar previsões ou classificações com maior precisão. Em estudos de ligas metálicas, por exemplo, as características derivadas podem incluir propriedades mecânicas, térmicas, elétricas, estrutura cristalina ou parâmetros de processamento [43,44].

Seleção do modelo de ML: A seleção do modelo depende da natureza do problema e dos dados utilizados. Diferentes algoritmos podem ser mais eficazes para diferentes cenários. Algoritmos comuns incluem máquinas de vetores de suporte (*Support Vector Machines* - SVM), redes neurais artificiais (*Artificial Neural Networks* - ANN), árvores de decisão (*Decision Trees*) e florestas aleatórias (*Random Forests* - RF), entre outros. Cada um desses modelos apresenta características específicas que os tornam mais adequados para determinados tipos de dados ou problemas [43,44,47].

Treinamento do modelo: O treinamento consiste no ajuste dos parâmetros do modelo utilizando um conjunto de dados de treinamento. Durante esse processo, o modelo aprende a mapear entradas (como a composição de uma liga metálica) para saídas desejadas (como dureza ou resistência). Esse passo é crucial para que o modelo capture as relações subjacentes entre as variáveis e forneça previsões precisas [43,44,47].

Validação do modelo: O desempenho do modelo é avaliado com base em um conjunto de dados de validação, distinto do conjunto de treinamento. Um método amplamente utilizado é a validação cruzada *K-fold*, onde os dados são divididos aleatoriamente em K partes. Em cada rodada, K-1 partes são utilizadas para treinamento, enquanto a parte restante é usada para validação. As métricas de avaliação variam conforme a tarefa. Em problemas de regressão, cujo objetivo é prever valores contínuos (como a dureza de uma liga metálica), o erro quadrático médio (*Root Mean Squared Error* - RMSE) é comumente utilizado. Em problemas de classificação, onde o objetivo é categorizar os dados em classes discretas (como identificar uma fase cristalina em uma liga), métricas como acurácia e pontuação F1 são empregadas para avaliar o desempenho do modelo [43,44].

3.4.1 Support Vector Machine

As máquinas de vetores de suporte (*Support Vector Machine* - SVM) é um método de aprendizado de máquina supervisionado amplamente utilizado em tarefas de

classificação. Introduzido por Boser, Guyon e Vapnik em 1992, ganhou destaque por sua robustez, boa capacidade de generalização e solução global ótima [47].

O objetivo principal do SVM é encontrar um hiperplano ótimo que separe duas classes distintas, maximizando a margem entre o hiperplano e os pontos mais próximos de cada classe. Esses pontos, conhecidos como vetores de suporte, são determinantes na definição da solução do modelo. O SVM se destaca por ser uma técnica esparsa, uma vez que, após o treinamento, depende apenas desses vetores de suporte para realizar previsões futuras [47,48]. O desenho esquemático dessa técnica pode ser visto na Figura 9.

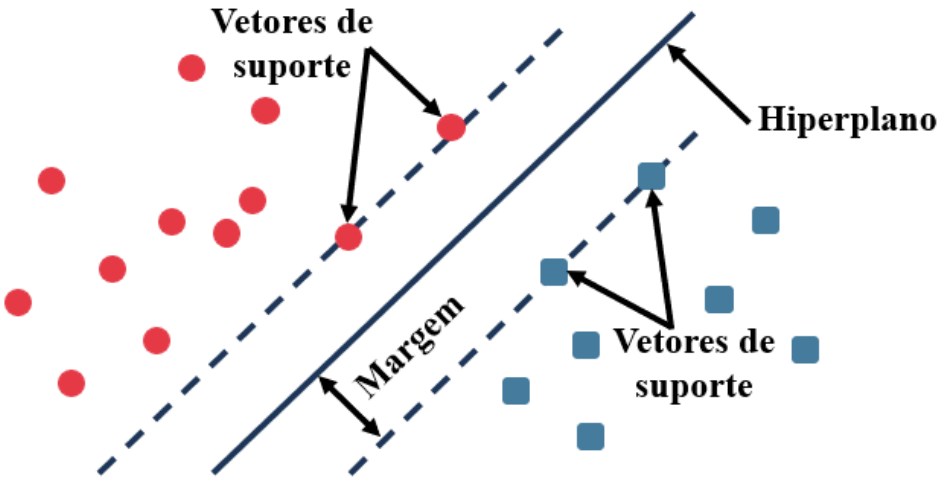

Figura 9 - Esquemática da técnica de *Support Vector Machine*. Destaca-se a presença de um hiperplano (linha preta contínua) separando dois grupos de dados (pontos vermelhos e verdes). A distância entre o hiperplano e os primeiros pontos de cada grupo (conhecidos como vetores de suporte) é denominada margem (linha preta tracejada).

Quando os dados não são linearmente separáveis no espaço original, o SVM utiliza uma técnica conhecida como *kernel trick*, que mapeia os dados para um espaço de características de maior dimensionalidade, onde a separação linear torna-se viável. Essa transformação é realizada por meio de funções de kernel, como polinomial, sigmoide, radial, base radial exponencial e linear, que permitem determinar uma fronteira de decisão não linear no espaço original [45,48,49]. Um exemplo pode ser conferido na Figura 10, na qual os pontos em um espaço 1D são elevados a um espaço 2D para serem separados linearmente.

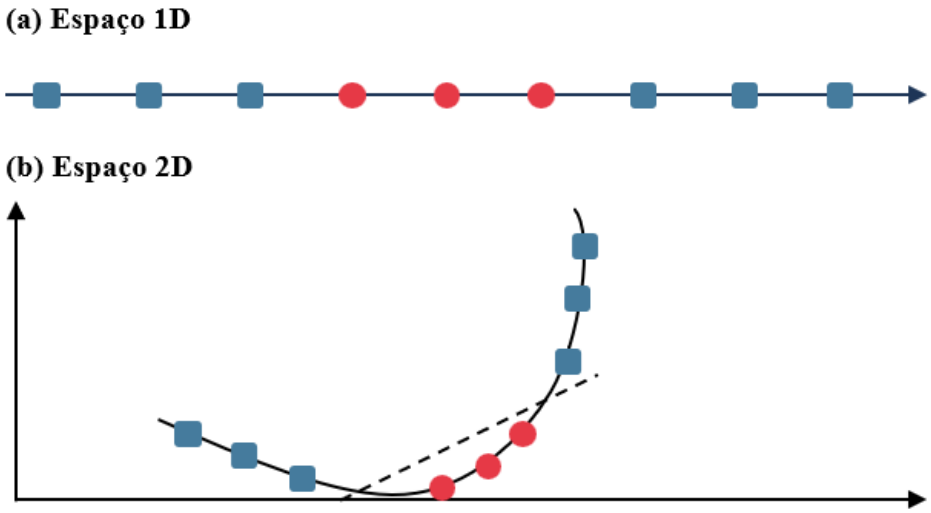

Figura 10 - Separação de dados por meio de função de Kernel. (a) Dois grupos de dados (pontos verdes e vermelhos) estão dispostos em um plano 1D sem possibilidade de separação linear. (b) Os dados são elevados em plano de dimensão superior (2D) e passam a ser passíveis de separação linear por meio do traço pontilhado.

Os hiperparâmetros do SVM exercem um papel crucial na eficácia do modelo e precisam ser cuidadosamente ajustados para garantir uma performance ótima. Os principais hiperparâmetros incluem:

1. *Kernel* e seus parâmetros associados:

- Tipo de *Kernel*:
    - *Kernel* Linear: gera uma fronteira de decisão reta, sendo adequado para dados separáveis linearmente [46,49].
    - *Kernel* Polinomial: cria uma fronteira curva e é frequentemente utilizado em problemas de processamento de imagens [46,49].
    - *Kernel* de Base Radial (RBF): é amplamente aplicado para dados não linearmente separáveis e apresenta um desempenho geral elevado [46,49].
    - *Kernel* Sigmoide: é usado em redes neurais como uma função de ativação, mas também pode ser aplicado no SVM [46,49].
- *Gamma*: utilizado nos *kernels* RBF, polinomial e sigmoide, define o grau de curvatura da fronteira de decisão. Valores altos de *gamma* levam a um contorno mais curvado, considerando apenas pontos próximos à fronteira, o que pode resultar em *overfitting*. Valores baixos consideram pontos mais distantes, proporcionando um contorno mais suave, mas com maior risco de *underfitting* [46,49].

- *Degree***:** aplicável ao *kernel* polinomial, determina o grau do polinômio utilizado e, consequentemente, a flexibilidade da fronteira de decisão [46,49].

2. Parâmetro de Regularização (*C*): Esse parâmetro regula o *trade-off* entre a maximização da margem e a minimização dos erros de classificação nos dados de treinamento. Valores altos de *C* priorizam a classificação correta dos dados de treinamento, mas podem levar a modelos complexos e com maior risco de *overfitting*. Por outro lado, valores baixos resultam em um modelo mais simples, com menor risco de *overfitting*, mas podem causar *underfitting* devido à permissão de um maior número de erros de classificação [46,47,49].

O SVM tem sido extensivamente aplicado na ciência dos materiais, destacando-se em estudos preditivos de formação de fases em ligas complexas. Vishwakarma e Neigapula [49] utilizaram dois modelos de SVM para prever a formação de fases em ligas de alta entropia. Um dos modelos baseou-se em dados de composição química, enquanto o outro utilizou características termodinâmicas. A base de dados foi construída a partir de informações experimentais publicadas, e os modelos alcançaram acurácia de aproximadamente 86%.

Bansal et al. [50] empregaram um modelo de SVM com *kernel* polinomial para identificar a formação de fases em HEAs para uma determinada temperatura. A base de dados foi gerada por cálculos CALPHAD de alto rendimento, abrangendo sistemas ternários, quaternários e quinários compostos pelos elementos Al, Fe, Co, Cr, Mn, Ni e V. O modelo apresentou uma acurácia de 99%, demonstrando alta capacidade preditiva para sistemas complexos.

Swateelagna et al. [51] focaram na predição de formação de fases em ligas refratárias de alta entropia (*Refractory High Entropy Alloys* - RHEAs) e na identificação dos fatores que afetam a precisão de modelos de aprendizado de máquina treinados. O modelo de SVM desenvolvido alcançou uma acurácia de teste de 86,67%, reforçando o potencial dessa abordagem em aplicações avançadas de ciência dos materiais.

3.4.2 Support Vector Regression

Support Vector Regression (SVR) é um método de aprendizado supervisionado baseado nos princípios do *Support Vector Machine* (SVM), adaptado para resolver problemas de regressão, ou seja, prever valores contínuos. O objetivo principal do SVR

é identificar uma função entre variáveis de entrada e saída, minimizando o erro de previsão e controlando a complexidade do modelo [47].

A transição de SVM para SVR é realizada introduzindo uma região ε-insensível ao redor da função de regressão, conhecida como tubo ε. Dentro desse tubo, erros absolutos inferiores a ε são ignorados, enquanto pontos fora do tubo são penalizados. Essa abordagem busca encontrar o tubo mais estreito que contenha a maioria dos dados de treinamento, equilibrando a complexidade do modelo e o erro de previsão. Em termos matemáticos, a formulação do SVR envolve a definição de uma função de perda ε-insensível que é minimizada durante a otimização. Assim, o SVR resolve essencialmente um problema de otimização multiobjetivo, que visa minimizar o erro de previsão e a complexidade geométrica do tubo [47].

No SVR, é possível trabalhar com dados não lineares ao mapear os dados de entrada para um espaço de alta dimensionalidade, utilizando funções kernel. Isso permite ao modelo capturar relações complexas entre as variáveis, mantendo uma capacidade robusta de generalização. Entre as funções kernel mais utilizadas estão o kernel linear, o RBF e o polinomial [52]. A representação esquemática do modelo de SVR pode ser conferido na Figura 11.

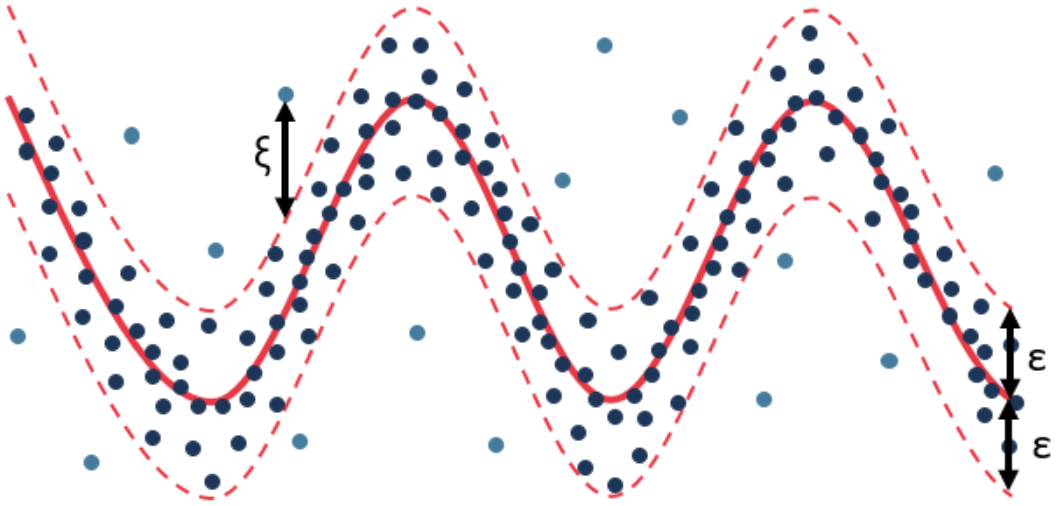

Figura 11 - Esquemática da técnica de *Support Vector Regression*. Destaca-se a presença da função que interliga os dados de entrada e saída por uma linha contínua, a medida em que o tubo de raio ε responsável por formar a região ε-insensível é representado por linhas tracejadas. Os pontos no interior do tubo são ignorados, enquanto os pontos no exterior do tudo a uma distância ξ são penalizados.

Os hiperparâmetros do SVR são fundamentais para determinar o desempenho do modelo e incluem:

- *C*: controla a penalidade por erros. Valores usuais variam de 0,001 a 1000. Valores mais altos priorizam a minimização do erro, mas podem levar ao *overfitting*. Valores mais baixos favorecem a generalização [52];
- *Epsilon* (ε): define a largura do tubo ε-insensível. Os valores comuns estão entre 0 e 1. Valores maiores tornam o modelo mais tolerante a desvios, enquanto valores menores aumentam a sensibilidade aos dados [52];
- *Kernel*: especifica o tipo de transformação dos dados. *Kernels* comuns incluem linear, RBF e polinomial [52];
- *Gamma* (γ): determina a influência de amostras individuais no modelo. Os valores usuais variam de 0,001 a 1000. Valores altos podem levar ao *overfitting*, enquanto valores baixos promovem a generalização [52].

Exemplos de aplicação do SVR destacam sua versatilidade e precisão. Em estudos sobre ligas refratárias de alta entropia, o SVR foi utilizado para prever a dureza e ductilidade de composições não exploradas, apresentando ótimos resultados frente a outros modelos de ML com RMSE de 10,3 HV para dureza e 3,58% para ductilidade [53]. Em outro exemplo, modelos de aprendizado de máquina, incluindo SVR, foram empregados para prever o módulo de elasticidade de ligas de alta entropia, sendo o SVR um dos três melhores, alcançando um coeficiente de determinação (R^2) de 92,4% e um RMSE de 4,4% [54]. Esses resultados demonstram a eficácia do SVR em prever propriedades físicas de materiais a partir de dados experimentais.

3.4.3 Active Learning

O aprendizado ativo (*active learning*) representa uma classe de métodos supervisionados de aprendizado de máquina focados em melhorar a acurácia dos modelos enquanto minimizam a quantidade de dados rotulados necessários para o treinamento [55,56]. Essa abordagem oferece diversas vantagens, como a aceleração do treinamento dos modelos, a redução da quantidade de dados de treinamento necessários, o encurtamento da duração do treinamento e o uso da menor quantidade possível de amostras rotuladas para alcançar alta precisão na classificação [57]. Essas características tornam o aprendizado ativo particularmente relevante em áreas como a ciência dos materiais, onde a obtenção de dados rotulados de alta qualidade frequentemente implica altos custos experimentais e computacionais [55].

A característica central do aprendizado ativo é sua capacidade de selecionar iterativamente instâncias não rotuladas que são consideradas mais informativas para o modelo. Essas instâncias são escolhidas com base em funções de utilidade ou aquisição, que avaliam quais pontos do espaço de busca possuem maior potencial de contribuir para a melhoria do desempenho do modelo preditivo [55,58].

Estratégias de aprendizado ativo podem ser classificadas em dois enfoques principais: *exploration* e *exploitation*. Na *exploration*, os exemplos não rotulados são selecionados em regiões do espaço de entrada ainda não amostradas. Esse processo busca cobrir uniformemente o espaço de dados, reduzindo áreas em que o modelo preditivo pode cometer erros. Por outro lado, na *exploitation*, os exemplos escolhidos estão em regiões já densamente amostradas, com o objetivo de refinar localmente as predições do modelo [58].

3.5 Algoritmo Genético

O algoritmo genético (*Genetic Algorithm* - GA) é uma classe de métodos de busca e otimização inspirada no processo de evolução natural descrito por Charles Darwin. O princípio fundamental por trás do GA é a ideia de "sobrevivência do mais apto", em que as soluções mais adaptadas ao problema em questão têm maior chance de serem selecionadas para gerar novas soluções. O GA é amplamente utilizado para resolver problemas complexos de otimização, especialmente aqueles que envolvem múltiplos objetivos, onde não existe uma única solução ótima e onde a exploração de um grande espaço de soluções é necessária [59,60].

O GA é um algoritmo baseado em população, o que significa que, em vez de explorar uma única solução de cada vez, ele trabalha com um conjunto de soluções candidatas, chamadas de indivíduos ou cromossomos. A cada geração, os indivíduos da população são avaliados com base em sua adequação à função de *fitness*, que quantifica sua capacidade de resolver o problema. A partir dessa avaliação, os melhores indivíduos são selecionados para gerar novos indivíduos por meio de operadores genéticos, como seleção, cruzamento e mutação. Esses operadores permitem que novas soluções sejam criadas e introduzidas na população, aumentando a diversidade e permitindo a exploração de diferentes regiões do espaço de soluções [59,61–63].

A principal característica dos GAs é sua natureza iterativa, em que, a cada ciclo, uma nova população de soluções é gerada, com o objetivo de melhorar as soluções existentes. Esse processo de evolução pode ser repetido várias vezes até que uma condição de convergência seja alcançada, como um número máximo de gerações ou uma melhoria insatisfatória nas soluções. Os GAs são, portanto, considerados métodos metaheurísticos, ou seja, estratégias de busca que são eficientes em encontrar soluções boas o suficiente para problemas difíceis de serem resolvidos de maneira exata, especialmente dentro de tempos computacionais razoáveis [60,64].

O método é constituído fundamentalmente de cinco etapas distintas, que serão descritas a seguir. Também é possível conferir um fluxograma simplificado do algoritmo na Figura 12.

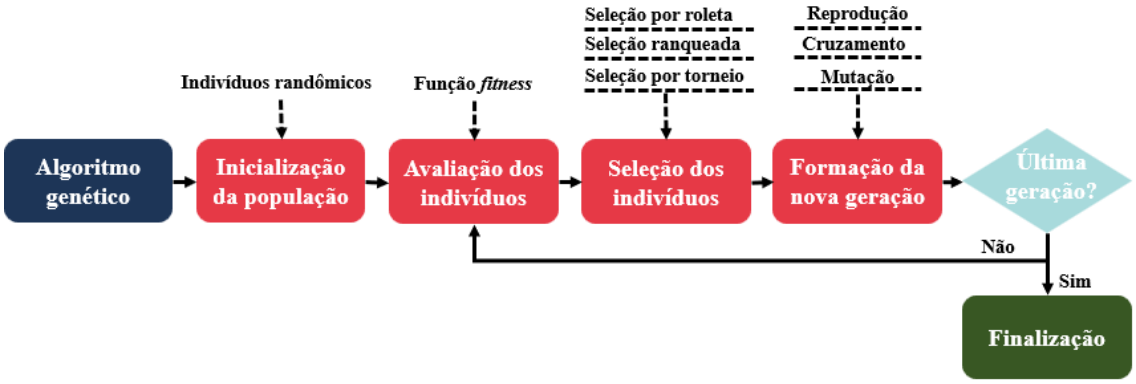

Figura 12 - Fluxograma das etapas de um algoritmo genético.

Inicialização da população: A inicialização da população é uma etapa essencial nos algoritmos genéticos, influenciando diretamente a qualidade das soluções. Nessa fase, uma população inicial de indivíduos é gerada para servir como ponto de partida. Geralmente, esses indivíduos são criados aleatoriamente, promovendo diversidade e permitindo a exploração do espaço de busca. Contudo, muitos desses indivíduos podem ser soluções inadequadas, enquanto outros, por acaso, podem estar mais próximos de uma solução viável [59,61,63].

O tamanho da população inicial é um parâmetro crítico. Populações maiores aumentam a diversidade e as chances de explorar regiões promissoras, mas demandam maior esforço computacional. Por outro lado, populações menores podem acelerar o processamento, mas correm o risco de gerar soluções subótimas. Assim, é necessário equilibrar diversidade e custo computacional de acordo com a natureza do problema [59].

Avaliação dos indivíduos: A avaliação dos indivíduos é um passo crucial nos algoritmos genéticos, pois determina a qualidade de cada indivíduo em relação ao problema a ser resolvido. Essa avaliação é conduzida por meio de uma função de aptidão, denominada *fitness function*, que atribui a cada indivíduo um valor numérico representando o quão bem ele atende ao objetivo do problema. Esse valor é usado para orientar o processo de seleção, direcionando a busca para as regiões mais promissoras do espaço de soluções [4,59,62,63].

Seleção dos indivíduos: A seleção é uma etapa fundamental nos algoritmos genéticos, sendo responsável por determinar quais indivíduos da população atual participarão do processo de reprodução para gerar a próxima geração. Essa etapa é baseada nos valores de *fitness* de cada indivíduo, atribuindo maior probabilidade de seleção àqueles com melhor desempenho no problema em questão [59].

Existem diversos métodos de seleção amplamente utilizados na literatura, cada um com características específicas e aplicações apropriadas. O método roleta (*roulette wheel*) é um dos mais conhecidos, onde os indivíduos são mapeados em uma roda proporcional aos seus valores de *fitness*. A seleção ocorre através de um giro aleatório da roda, garantindo que soluções mais aptas tenham maior chance de participar da próxima geração. Uma variação desse método é a seleção por rank, que utiliza as classificações dos indivíduos em vez de seus valores absolutos de *fitness*, reduzindo o risco de convergência prematura para mínimos locais [59,62,65,66].

Outro método amplamente empregado é a seleção por torneio. Nesse caso, pares de indivíduos são selecionados de forma estocástica, e o indivíduo com melhor *fitness* é adicionado ao *pool* de reprodução. Essa técnica é conhecida por oferecer a todos os indivíduos uma chance de serem selecionados para participarem no processo de cruzamento para formação de uma nova geração [59,65,66].

Além disso, outras estratégias, como seleção truncada, seleção por amostragem universal estocástica, seleção Boltzmann, dentre outras, também são utilizadas dependendo das características do problema e das restrições computacionais. Cada uma dessas técnicas busca equilibrar a pressão de seleção. Ela é definida como o grau em que os indivíduos mais aptos são favorecidos no processo de seleção. Pressões de seleção muito altas podem acelerar a convergência, mas também aumentam o risco de estagnação prematura em soluções subótimas. Por outro lado, pressões muito baixas podem

desacelerar a convergência, levando o algoritmo a um tempo excessivo para encontrar soluções otimizadas. O equilíbrio adequado da pressão de seleção é, portanto, essencial para o desempenho do GA, que influencia diretamente a taxa de convergência e a capacidade de escapar de soluções subótimas [62,65,66].

Formação da nova geração: A formação de uma nova população em algoritmos genéticos é realizada por meio de três operadores genéticos principais, a exceção da já apresentada seleção: reprodução, *crossover* e mutação. Esses operadores são fundamentais para garantir a exploração do espaço de busca e evitar a convergência prematura a soluções subótimas.

Reprodução

O operador de reprodução é o responsável por garantir que os melhores indivíduos de uma população sejam preservados diretamente na geração seguinte sem modificações. Esse processo assegura que as melhores soluções encontradas até o momento não sejam perdidas devido à aplicação de operadores que podem introduzir variações desfavoráveis [59,66].

Cruzamento

O cruzamento, ou *crossover*, combina informações genéticas de dois ou mais indivíduos (pais) para criar novos indivíduos (filhos), simulando a troca de material genético na reprodução sexual. Dentre as técnicas disponíveis, citam-se o *crossover* de ponto único, multiponto, uniforme, parcialmente mapeado (PMX), e baseado em ordem ou posição. O método mais adequado depende do problema em questão e da representação dos cromossomos [59,62,65,66].

No *crossover* de ponto único (Figura 13.a), um ponto de corte é escolhido aleatoriamente, e os segmentos dos pais após esse ponto são trocados. No *crossover* multiponto (Figura 13.b), vários pontos de corte são selecionados, permitindo uma troca mais detalhada de segmentos genéticos. Já no *crossover* uniforme (Figura 13.c), cada gene é tratado individualmente, com a decisão de troca sendo feita aleatoriamente para cada posição do cromossomo [59,62,65,66].

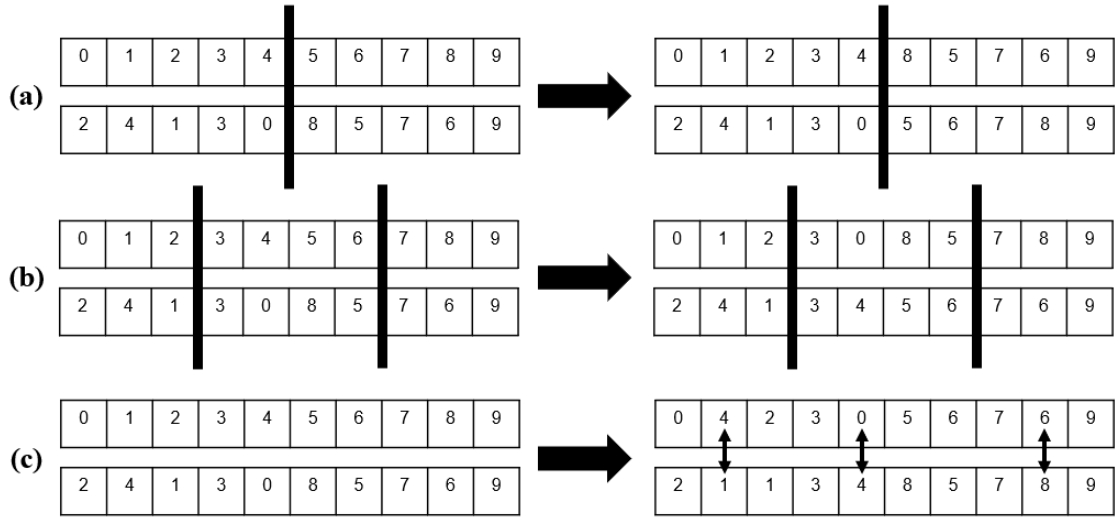

Figura 13 - Operador genético *crossover*: (a) *crossover* de ponto único, (b) *crossover* multiponto e (c) *crossover* uniforme.

Mutação

Após o cruzamento, a mutação é aplicada para introduzir pequenas alterações nos genes dos indivíduos, garantindo a diversidade genética da população e prevenindo que todas as soluções se fixem em um ótimo local. A mutação é especialmente importante em problemas onde há risco de estagnação, pois pode explorar novas áreas do espaço de busca [59,62,64].

A taxa de mutação (probabilidade de mutação) deve ser cuidadosamente ajustada. Taxas muito altas transformam o GA em uma busca aleatória, prejudicando o impacto do *crossover* e da seleção. Por outro lado, taxas muito baixas limitam a capacidade do algoritmo de escapar de ótimos locais, reduzindo a eficiência geral [59,62,64].

Finalização do algoritmo genético: A finalização de algoritmos genéticos ocorre quando critérios específicos são atendidos. Os critérios mais comuns incluem a obtenção de uma solução que atenda aos requisitos mínimos, o alcance de um número máximo de gerações, o uso completo do orçamento computacional disponível ou ainda a estabilização da função *fitness* da melhor solução encontrada, indicando que não há melhorias significativas nas gerações subsequentes, o que evita o uso excessivo de recursos. Esses critérios podem ser aplicados de forma individual ou combinada, dependendo do problema e dos objetivos do algoritmo [60,63].

Os algoritmos genéticos têm se destacado em diversas aplicações na engenharia de materiais, especialmente em problemas que envolvem otimização de propriedades complexas. Um exemplo relevante é a otimização de ligas de alta entropia, onde um

algoritmo genético modificado foi utilizado para explorar o vasto espaço de busca dessas ligas. Neste caso, o objetivo foi maximizar a estabilidade de soluções sólidas de estrutura CFC, melhorar o endurecimento por solução sólida e minimizar a densidade. Essa abordagem demonstrou a eficácia dos GAs em superar desafios associados à complexidade das composições e aos múltiplos critérios de projeto [67].

Outro uso significativo dos algoritmos genéticos foi no design de superligas à base de níquel, como nos sistemas Ni–Al–Cr–Mo–Ta e Ni–Al–Cr–Co–W–Ti–Ta. Nessa aplicação, foram utilizadas simulações de dinâmica molecular em conjunto com GAs para prever composições ideais com base em um grande número de parâmetros físicos, abordando sistemas multicomponentes complexos [68].

Adicionalmente, um estudo baseado em uma abordagem evolutiva utilizou algoritmos genéticos para identificar as 20 ligas mais estáveis entre 192.016 configurações possíveis de 32 metais diferentes em estruturas CFC e CCC. Nesse caso, a entalpia de formação foi utilizada como critério de estabilidade, com cálculos baseados na teoria do funcional da densidade. O uso de GAs permitiu uma busca eficiente em um espaço extenso, evidenciando seu potencial no desenvolvimento de novos materiais [69].

Dessa forma, os algoritmos genéticos se destacam como ferramentas poderosas para resolver problemas complexos de otimização em engenharia de materiais. Suas principais vantagens incluem a capacidade de explorar amplos espaços de busca, identificar soluções globais em problemas multimodais e lidar com funções ruidosas ou descontínuas. Além disso, a adaptabilidade dos GAs permite sua aplicação em uma ampla gama de problemas. Contudo, os desafios relacionados à escolha de parâmetros, configuração inicial e possível elevado custo computacional devem ser considerados com atenção [66].

4 MATERIAIS E MÉTODOS

Os materiais e métodos adotados neste projeto estão resumidos na Figura 14. Os códigos desenvolvidos neste trabalho foram executados em uma estação de trabalho equipada com processador Intel® Xeon® Silver 4214R (2,40 GHz × 48 núcleos), memória RAM de 156,8 GiB e placa gráfica NVIDIA Quadro P1000. Nos tópicos subsequentes, cada etapa será detalhada, com ênfase nos procedimentos adotados e nas ferramentas empregadas. Todos os códigos empregados na metodologia podem ser acessados em https://github.com/Caroline-B-S/IA-for-HEA-design.

Figura 14 - Fluxograma dos materiais e métodos a serem utilizados no trabalho.

4.1 Escolha do espaço composicional

Os elementos químicos considerados para a formulação das ligas investigadas neste trabalho foram alumínio (Al), cobalto (Co), cromo (Cr), ferro (Fe), manganês (Mn), níquel (Ni) e vanádio (V). A escolha foi fundamentada em critérios científicos e práticos, incluindo: (i) a ampla disponibilidade de dados na literatura acerca das propriedades desses elementos, com ênfase nos valores de energia de falha de empilhamento, essenciais para a previsão computacional do comportamento mecânico das ligas; (ii) a inclusão desses elementos na base de dados PanHEA2023, utilizada como base termodinâmica na previsão de fases das ligas de alta entropia pelo método CALPHAD; e (iii) a disponibilidade física dos elementos nos laboratórios, assegurando a viabilidade prática para futura síntese e caracterização experimental.

4.2 Criação das bases de dados

4.2.1 Base de dados para previsão da fase

A obtenção dos dados brutos para previsão de fases em ligas de alta entropia utilizados neste trabalho foi realizada por meio do software Pandat™, desenvolvido pela CompuTherm LLC, empresa que produz soluções para cálculos termodinâmicos e de diagramas de fase baseada na abordagem CALPHAD. O Pandat™ conta com diversos módulos, incluindo o PanPhaseDiagram, projetado para calcular equilíbrios de fases e propriedades termodinâmicas em sistemas multicomponentes e multifásicos. Este módulo permite determinar diagramas de fases estáveis e metaestáveis, temperaturas de transformação de fase, frações de fases e propriedades termodinâmicas, como energia livre de Gibbs, entalpia e entropia, além de realizar simulações de solidificação com base no modelo de Scheil [70].

Adicionalmente, foi empregado o PanPython SDK (*Software Development Kit* - SDK), uma interface que permite a integração dos módulos do Pandat™ ao ambiente de programação Python, otimizando o fluxo de trabalho ao possibilitar o uso de bibliotecas numéricas, estatísticas e de análise de dados amplamente disponíveis. O PanPython oferece recursos específicos para cálculos de alto rendimento (*High-Throughput Calculation* - HTC), permitindo a realização de cálculos em paralelo sob condições definidas pelo usuário. Essa funcionalidade foi explorada neste trabalho para gerar dados em larga escala de maneira eficiente e assertiva [71].

Com base nas condições definidas pelo usuário, foram geradas 59.710 composições distintas, representando uma malha completa de pontos formada pelas combinações dos sete elementos Al, Co, Cr, Fe, Mn, Ni e V. As concentrações desses elementos foram variadas entre 0% e 30% atômico, com incrementos de 5% atômico, respeitando a restrição de que a soma total das concentrações não excedesse 100%. O níquel foi utilizado como elemento balanceador, completando o total de 100% nas composições. A seleção desses elementos, já justificada previamente, considera tanto a disponibilidade de dados na literatura, quanto a disponibilidade física do material, bem como a sua presença na base de dados PanHEA2023, utilizada como base termodinâmica nos cálculos realizados pelo Pandat™. A escolha da faixa de concentração de 0 a 30% atômico para cada elemento está em conformidade com a definição de ligas de alta entropia, nas quais a concentração de um único elemento deve ser inferior a 35% [1,8,9].

O passo de 5% atômico foi determinado de maneira a equilibrar a precisão da malha com a eficiência computacional, permitindo uma representação adequada das possíveis fases formadas sem comprometer excessivamente o custo computacional. Para cada composição, foram calculadas as fases resultantes a uma temperatura de 900°C, escolhida por ser suficientemente alta para a estabilização de possíveis ligas monofásicas CFC, mas ainda dentro de uma temperatura onde a presença de fase líquida é suprimida ou minimizada. Inicialmente, foi construída uma base de dados a 1000 °C; no entanto, verificou-se que, nessa temperatura, mais da metade das composições apresentava fase líquida como segunda fase. Esse comportamento dificultava a análise voltada exclusivamente para ligas completamente sólidas. Dessa forma, optou-se por realizar os cálculos a 900 °C, o que permitiu a construção de uma base de dados mais adequada ao objetivo do trabalho.

Para a construção da base de dados final, foi realizado um filtro inicial para selecionar apenas as composições cuja descrição incluísse um número adequado de binários e ternários presentes na base termodinâmica (PanHEA2023). Inicialmente, a descrição completa da base PanHEA2023 foi analisada, identificando os binários e ternários ali descritos. Em seguida, para cada composição gerada pelo PanPython, foram extraídos todos os binários e ternários que a compunham. Observou-se que todos os binários estavam descritos na base PanHEA2023, de modo que o critério de filtro foi aplicado apenas em relação aos ternários descritos. Para determinar o limite adequado de filtragem, diferentes valores de FAT (*Fully Assessed Ternary systems*) foram testados, variando de 10% a 90%, em incrementos de 10%. O valor final selecionado foi aquele que representou o melhor compromisso entre a manutenção de um elevado número de ternários descritos e a minimização da perda de dados, garantindo a qualidade e representatividade da base de dados resultante. A relação entre o número mínimo de ternários descritos na base de dados PanHEA2023 e a porcentagem de dados removidos é apresentada na Figura 15.

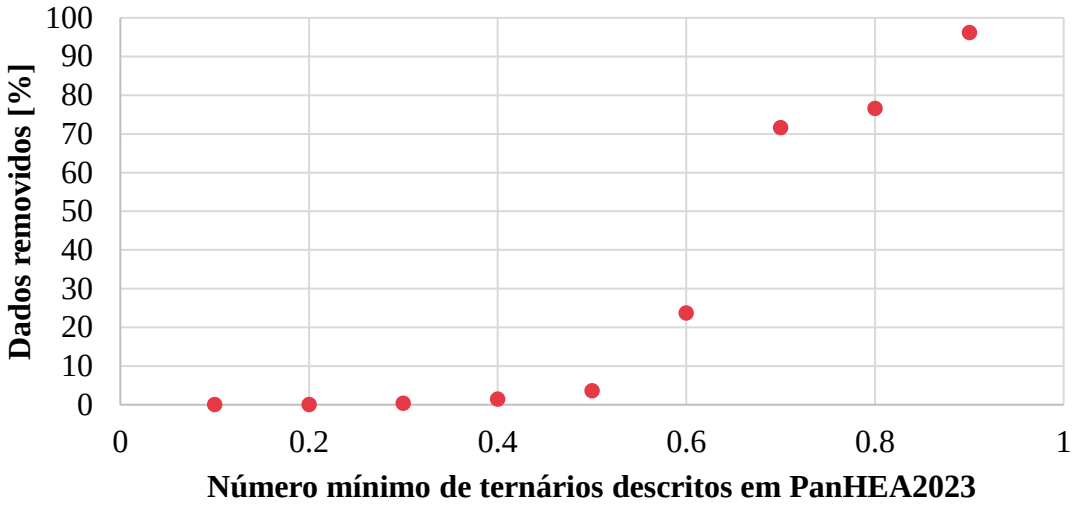

Figura 15 - Relação entre o número mínimo de ternários descritos na base de dados PanHEA2023 e a porcentagem de dados removidos.

Com base nessa análise, optou-se por adotar um valor de FAT igual a 60% como critério de filtragem para a formação da nova base de dados. Essa escolha resultou na exclusão de 23.74% das composições inicialmente geradas, representando um compromisso adequado entre a qualidade da base de dados e a preservação de um volume significativo de informações. Valores de FAT superiores a 60% resultaram na remoção de mais de 71.61% dos dados (como observado com FAT de 70%) e eliminaram completamente todas as composições contendo os sete elementos considerados. Tal exclusão seria prejudicial, uma vez que essas composições representam combinações-chave para a análise proposta e foram fundamentais na construção da malha inicial de composições via PanPython. Ademais, a escolha de um valor de FAT igual a 60% está em consonância com estudos previamente publicados na literatura [72], sendo inclusive considerado um valor conservador.

Após o processo de filtragem, a base de dados foi aprimorada para garantir maior equilíbrio entre as classes de composições. Dentre as 45.535 composições resultantes após filtragem, observou-se que 5.33% eram monofásicas CFC, 49.92% apresentavam a fase CFC combinada com outra fase, e 44.75% consistiam exclusivamente de outras fases que não CFC. Essa distribuição evidenciou um desbalanceamento significativo na base de dados, especialmente para a classe de ligas monofásicas CFC, as quais são de essencial importância neste trabalho. Para minimizar esse problema, novos pontos foram adicionados de acordo com a seguinte lógica, a qual se encontra ilustrada de modo simplificado para dois elementos na Figura 16.

Inicialmente, para cada composição contendo CFC e uma segunda fase (Figura 16, círculo azul), foi extraída a composição correspondente da fase CFC utilizando o princípio adaptado da regra da alavanca (Figura 16, triângulo vermelho), informação fornecida diretamente pela tabela gerada no PanPython. Em seguida, considerou-se um vetor conectando a composição monofásica CFC derivada e a composição original contendo CFC mais uma segunda fase (Figura 16, linha tracejada preta). Ao longo desse vetor, foi selecionado um único outro ponto adicional a uma distância fixa de 10% da composição monofásica CFC (Figura 16, quadrado amarelo), avançando em direção à composição original da base de dados. Essa distância foi definida com base em testes exploratórios, nos quais distâncias menores, como 2% e 5%, resultaram em composições com variações atômicas pequenas (frequentemente restritas à primeira ou segunda casa decimal), o que poderia comprometer a diversidade dos dados e a eficácia do treinamento do algoritmo. O valor de 10% mostrou-se adequado por introduzir uma diferença composicional suficientemente significativa, ao mesmo tempo em que mantém o novo ponto próximo à fronteira entre as regiões CFC e CFC + segunda fase, favorecendo a definição de um hiperplano mais robusto pelo classificador SVM. Essa estratégia permitiu explorar regiões intermediárias no espaço composicional de forma eficiente e controlada.

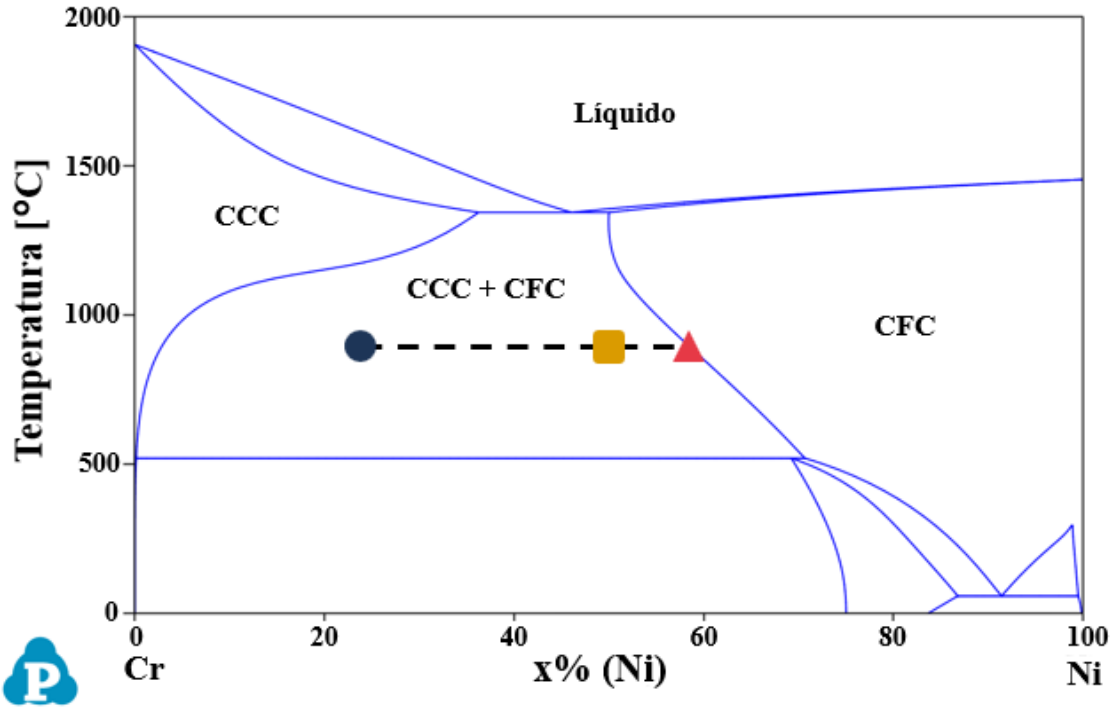

Figura 16 - Diagrama de fases Cr-Ni obtido utilizando o software Pandat™ e a base de dados PanHEA2023. Os símbolos representam diferentes composições no espaço composicional: o círculo azul indica a composição original da base de dados, o triângulo vermelho corresponde à composição monofásica CFC derivada pela aplicação da regra da alavanca, e o quadrado amarelo representa uma composição intermediária gerada entre os dois pontos anteriores.

Essa abordagem resultou na criação de uma base final contendo 80.553 composições distintas, representando um aumento de aproximadamente 77% em relação à base filtrada inicial. Após a adição desses novos pontos, a distribuição final das classes foi aprimorada, com 24.75% das composições sendo monofásicas CFC, 49.95% contendo CFC mais uma segunda fase, e 25.30% apresentando apenas outras fases.

Esse enriquecimento trouxe diversas vantagens: além de aumentar o número total de composições disponíveis para o treinamento do algoritmo de *machine learning*, proporcionou uma base de dados mais balanceada, essencial para melhorar a performance do modelo. No caso específico do algoritmo de SVM, os pontos adicionais são fundamentais para auxiliar na definição de vetores de suporte, contribuindo para a formação de um hiperplano mais preciso na separação entre as classes.

Por fim, considerando que o problema em questão é de classificação, as composições foram categorizadas de forma numérica para facilitar o treinamento do modelo de ML. As ligas monofásicas CFC foram designadas com o número 10, as composições contendo CFC mais uma segunda fase receberam o número 7, e as demais composições foram identificadas pelo número 0.

4.2.2 Base de dados para previsão da energia de falha de empilhamento

A base de dados utilizada para a previsão da energia de falha de empilhamento, fundamental para determinar as possíveis ativações dos efeitos TWIP/TRIP, foi obtida a partir do estudo de Khan et al. [33] apresentada na seção 3.2.3.2. Neste trabalho, os autores calcularam a EFE de 489 ligas compostas pelos elementos Al, Co, Cr, Fe, Mn, Ni e V por meio de DFT.

Vale-se ressaltar que a seleção dos elementos apresentados na Seção 3.1 foi diretamente influenciada pela base de dados de energia de falha de empilhamento utilizada e descrita por Khan et al. [33]. Optou-se por manter os mesmos elementos da base original (Al, Co, Cr, Fe, Mn, Ni e V), sem realizar cálculos adicionais por DFT ou extrapolações para outros elementos. Essa abordagem conservadora visou preservar a qualidade e a confiabilidade dos resultados, uma vez que não é possível prever com precisão o comportamento da base de dados em extrapolações para sistemas que incluam elementos diferentes. Isso se deve à característica intrínseca da base, na qual todas as ligas foram geradas como combinações contendo simultaneamente os sete elementos

mencionados. Consequentemente, não era viável separar um subconjunto da base para teste que fosse composto por ligas de ordens distintas, dificultando qualquer validação robusta de extrapolações.

4.3 Seleção de descritores

Como discutido na seção 2.4, a engenharia de características consiste em derivar propriedades relevantes a partir dos dados brutos, com o objetivo de representá-los de maneira mais adequada para a construção de modelos preditivos. Os descritores de materiais são parâmetros quantitativos que caracterizam as propriedades de um material específico. A eficácia e a capacidade preditiva dos modelos de *machine learning* dependem diretamente da seleção apropriada desses descritores, que devem representar o conjunto de dados de forma eficaz. O uso de descritores, além dos dados de composição bruta, possibilita extrapolações além dos limites iniciais do conjunto de dados, permitindo a previsão de novos materiais e propriedades não contemplados pelos dados de treinamento [33,73].

Assim, a transformação da composição bruta das bases de dados para previsão de fase e de energia de falha de empilhamento em propriedades específicas garante não apenas uma melhora na capacidade preditiva dos modelos de ML, mas também abre caminho para, em trabalhos futuros, extrapolar a base de dados e prever ligas contendo elementos diferentes dos sete inicialmente considerados.

No contexto da previsão de fases e de energia de falha de empilhamento, os descritores mais comumente utilizados abrangem uma ampla gama de propriedades, incluindo raio atômico, temperatura de fusão, eletronegatividade, concentração de elétrons de valência, entalpia de mistura, entropia de mistura e módulo de cisalhamento [33,67,74,75]. Esses descritores fornecem uma representação multidimensional dos materiais, permitindo que algoritmos de *machine learning* capturem relações complexas entre composição, estrutura e propriedades.

Neste estudo, as composições químicas presentes nas bases de dados utilizadas para a previsão de fases e energia de falha de empilhamento (seções 3.2.1 e 3.2.2) foram convertidas em 36 descritores. Os descritores empregados incluíram a média ponderada pela composição e o desvio padrão das seguintes propriedades: concentração de elétrons de valência; densidade; eletronegatividade; entalpia de fusão; entropia de fusão; grupo da

tabela periódica; massa atômica; módulo de cisalhamento; número atômico; número de Mendeleev (baseado em uma organização dos elementos químicos de acordo com a massa atômica dos seus átomos); período da tabela periódica; raio atômico; temperaturas de fusão e ebulição e volume molar. Foram utilizados também: desajuste de tamanho atômico; entalpia de mistura; entropia de mistura ideal; soma da entropia de mistura ideal com a entropia do elemento puro na estrutura CFC a 900 °C; capacidade térmica de excesso e soma da capacidade térmica de excesso com a capacidade térmica específica do elemento puro na estrutura CFC a 900 °C. Um resumo dos descritores com suas respectivas equações pode ser conferido no APÊNDICE A.

A entalpia de mistura e a capacidade térmica de excesso foram calculadas conforme o método descrito por Deffrennes et al. [76]. A entropia e a capacidade térmica específica dos elementos puros na estrutura CFC a 900 °C foram obtidas por meio do software PANDAT™ (base de dados PanHEA2023). O módulo de cisalhamento foi extraído da Element Collection, Inc. [77], a qual utilizou a função *ElementData* do *Mathematica*, desenvolvido pela Wolfram Research, Inc. As demais propriedades foram extraídas do trabalho de Ward et al. [78]. No Apêndice A é possível conferir as equações utilizadas no cálculo de cada um dos descritores mencionados.

A seleção dos descritores mais relevantes foi realizada utilizando a técnica de seleção sequencial progressiva (*Forward Sequential Selection* - FSS). Esse método visa reduzir a dimensionalidade do conjunto de dados, aprimorando tanto o desempenho quanto a interpretabilidade dos modelos. Algoritmos de busca sequencial utilizam uma estratégia de otimização do tipo *hill-climbing*, na qual características são iterativamente adicionadas ou removidas. Especificamente, a FSS inicia com um conjunto vazio de descritores e os adiciona de forma incremental até formar um subconjunto otimizado [79]. O processo é estruturado em três etapas principais:

1. **Seleção inicial**: o modelo é treinado e testado com cada descritor individualmente. A métrica de desempenho do modelo resultante é avaliada, e o descritor que apresentar o melhor resultado é selecionado [79]. A métrica utilizada depende do tipo de problema: no caso da classificação por SVM, foi considerada a acurácia; para regressão com SVR, foi utilizado o erro quadrático médio da raiz (RMSE);

2. **Adição iterativa**: características adicionais são incorporadas uma a uma, priorizando-se aquelas que mais contribuem para a melhoria do desempenho do modelo [79];
3. **Critério de parada**: o processo é encerrado quando o número desejado de características é alcançado ou quando não há mais ganho significativo no desempenho do modelo [79].

4.4 Otimização dos hiperparâmetros

4.4.1 Otimização dos hiperparâmetros para previsão da fase

O *Support Vector Machine* foi selecionado como o algoritmo de *machine learning* a ser utilizado na previsão das fases. A escolha foi motivada pelos resultados obtidos no trabalho de Santos [80]. Nesse estudo, foi comparada a performance de três algoritmos supervisionados – K-vizinhos mais próximos (KNN), árvore de decisão e SVM – na classificação de fases de ligas de alta entropia formadas por Co, Cr, Fe, Ni e Mn, com dados obtidos por meio de cálculos termodinâmicos via método CALPHAD. Com uma base inicial de 1.000 composições e validando 494 novas combinações, o SVM apresentou a maior acurácia (96.36%), seguido pelo KNN (94.95%) e pela árvore de decisão (91.72%). Além disso, o SVM demonstrou especial robustez na classificação de ligas monofásicas CFC, classificando corretamente 312 das 315 composições monofásicas CFC presentes na base de validação. A semelhança com os elementos trabalhados, a forma de obtenção da base de dados e o objetivo da classificação, consolidaram o SVM como a melhor escolha para este trabalho de mestrado.

Conforme detalhado na Seção 3.4, o algoritmo SVM apresenta diversos hiperparâmetros que podem ser ajustados para maximizar a acurácia do modelo. Entre os mais relevantes estão o parâmetro de regularização C, que controla o equilíbrio entre a complexidade do modelo e a penalização por erros de classificação, e as configurações associadas ao *Kernel*, como o tipo de *Kernel* a ser utilizado e os parâmetros *gamma* e *degree*.

O processo de ajuste dos hiperparâmetros começa com a definição de faixas de variação adequadas. Considerando a dimensão significativa da base de dados, composta por 80.553 composições, essas faixas foram estabelecidas visando um compromisso entre

a maximização da acurácia do modelo e a eficiência computacional, medida pelo tempo de processamento necessário.

Para tanto, a determinação de uma faixa adequada foi realizada através de uma análise sistemática, variando-se apenas um hiperparâmetro por vez, enquanto os demais eram mantidos fixos em seus valores padrão, conforme definidos pela biblioteca Scikit-learn (*C* = 1.0, *kernel* = 'rbf', *degree* = 3, *gamma* = 'scale') [81]. Essa abordagem permite avaliar individualmente o impacto de cada parâmetro na acurácia do modelo e no tempo computacional. Para todos os experimentos, 20% da base de dados foi reservada para o conjunto de teste, um valor frequentemente adotado pela comunidade científica [82].

Parâmetro de regularização *C*

O hiperparâmetro *C* foi testado em uma ampla faixa de valores (10^{-5} a 10^{6}) (Figura 17). Observou-se que valores muito baixos ($C \leq 0.001$) resultaram em baixa acurácia (<70%), enquanto valores intermediários (C = 0.1 a 1000) apresentaram o melhor compromisso entre desempenho e eficiência computacional, alcançando acurácias superiores a 88% com tempos de processamento abaixo de 450 segundos. Valores muito elevados de C ($\geq 10^{5}$) também proporcionaram alta acurácia, mas com um custo computacional significativamente maior, chegando a aproximadamente 60.900 segundos para $C = 10^{6}$. Dessa forma, optou-se por restringir a faixa de *C* entre 0,1 e 1.000.

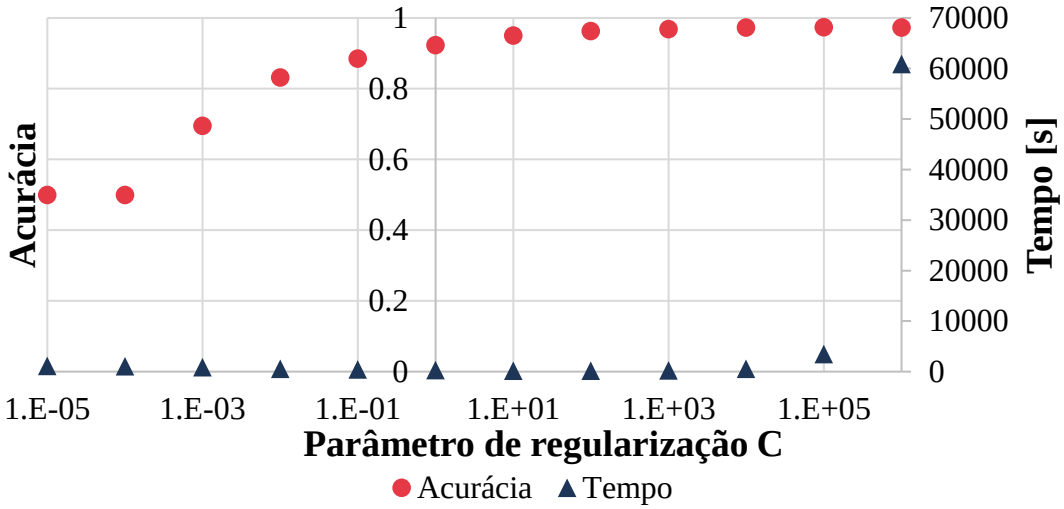

Figura 17 - Correlação entre possíveis valores de parâmetro de regularização C e seus respectivos valores de acurácia (círculos vermelhos) e tempo de processamento (triângulos azuis).

Tipo de *Kernel*

Foram avaliados quatro tipos de *kernel*: linear, polinomial (poly), *radial basis function* (rbf) e sigmoid (Figura 18). O *kernel* radial destacou-se com a maior acurácia (92.3%) e um tempo de execução moderado (320 segundos), enquanto o *kernel* polinomial apresentou o melhor tempo (210 segundos) e acurácia levemente inferior (90.3%). O *kernel* sigmoid demonstrou desempenho inferior aos demais, com acurácia de 50.7% e maior tempo de execução (437 segundos), sendo, portanto, eliminado das próximas análises.

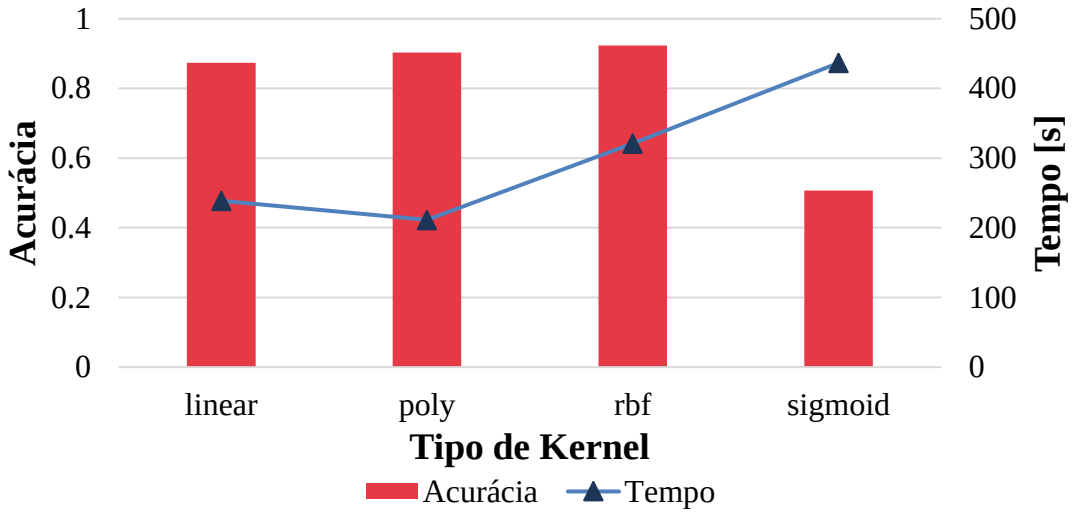

Figura 18 - Correlação entre possíveis tipos de Kernel e seus respectivos valores de acurácia (colunas vermelhas) e tempo de processamento (triângulos azuis).

Degree

O parâmetro *degree*, relevante para o *kernel* polinomial, foi variado de 2 a 10 (Figura 19). Verificou-se que valores mais baixos de *degree*, especialmente 3, proporcionaram melhor acurácia, cerca de 90.3%, e menores tempos de processamento. Valores mais altos resultaram em aumento do tempo de execução e redução da acurácia. Consequentemente, para configurações que utilizam o *kernel* polinomial, uma faixa de 2 a 5 foi escolhida.

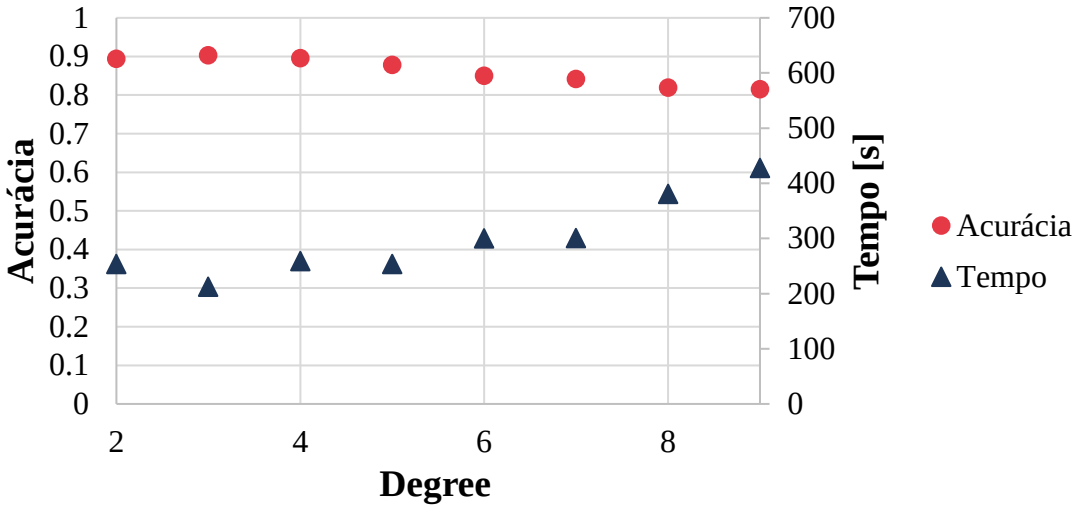

Figura 19 - Correlação entre possíveis valores de *degree* e seus respectivos valores de acurácia (círculos vermelhos) e tempo de processamento (triângulos azuis).

Gamma

As opções *auto* e *scale* foram testadas para o parâmetro *gamma* (Figura 20), ambas resultando em acurácias semelhantes (92.3%) e tempos de processamento próximos (315 segundos). Optou-se dessa forma por manter ambas as opções.

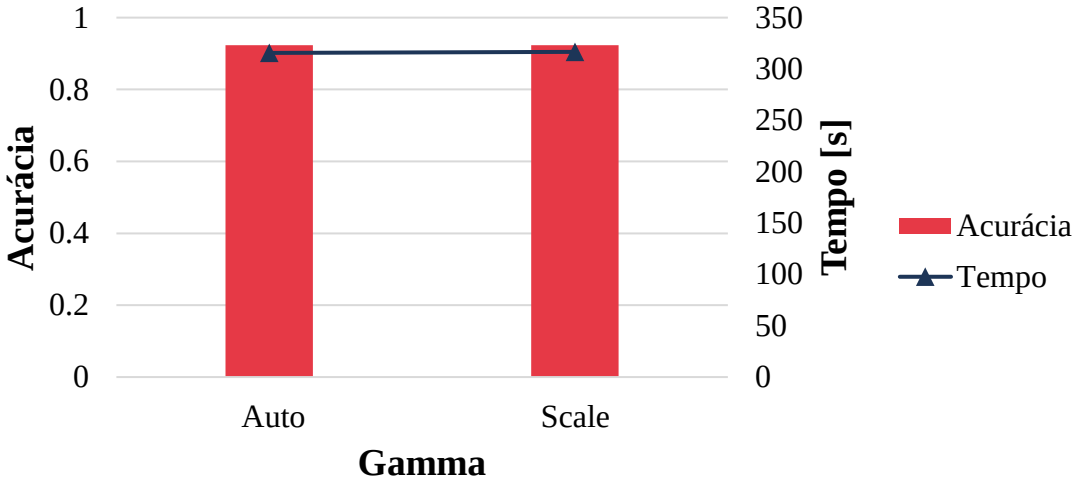

Figura 20 - Correlação entre possíveis tipos de *gamma* e seus respectivos valores de acurácia (colunas vermelhas) e tempo de processamento (triângulos azuis).

Um resumo dos principais hiperparâmetros dos SVM e suas respectivas faixas ou opções pode ser conferido na Tabela 2.

Tabela 2 - Resumo dos principais hiperparâmetros do modelo SVM e suas respectivas faixas ou opções.

Hiperparâmetro	Faixa / Opção
Parâmetro de regularização C	0.01 - 1000
Tipo de Kernel	rbf, poly, linear
Degree	2 - 5
Gama	auto, scale

Para a otimização simultânea desses múltiplos hiperparâmetros, foi utilizada a biblioteca Optuna, uma ferramenta de código aberto projetada para a otimização automatizada de hiperparâmetros [83]. Optuna permite explorar de forma eficiente espaços de busca amplos e complexos, integrando condicionais e *loops* em Python para implementar estratégias de busca dinâmicas e flexíveis. Além disso, a biblioteca utiliza técnicas de *pruning* (poda), que descartam tentativas pouco promissoras, acelerando o processo de busca e, consequentemente, aprimorando os resultados obtidos [83].

Na prática, a seleção dos descritores e a otimização dos hiperparâmetros foram realizadas de forma simultânea, como ilustrado na Figura 21. O processo seguiu uma abordagem iterativa, em que a cada iteração, um descritor era selecionado utilizando a técnica de seleção sequencial progressiva. Para cada conjunto de descritores formado, os hiperparâmetros eram otimizados via Optuna, e, com base na configuração resultante, um novo descritor era incorporado ao conjunto. Esse ciclo foi repetido até que todos os descritores fossem selecionados.

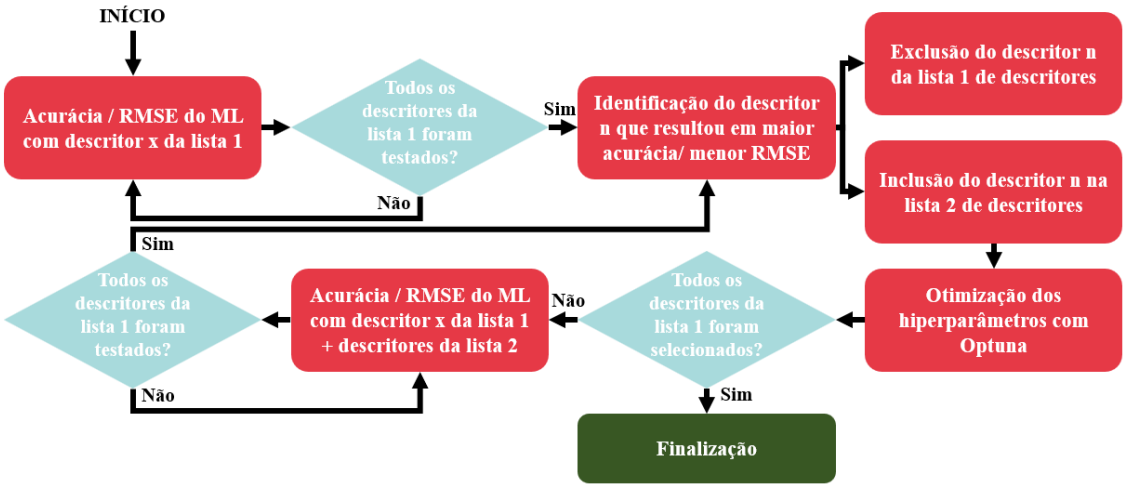

Figura 21 - Fluxograma de como foi realizado simultaneamente a escolha dos descritores e da otimização dos hiperparâmetros para o modelo SVM. A lista 1 se refere a descritores candidatos e a lista 2 se refere a descritores já testados e selecionados.

4.4.2 Otimização dos hiperparâmetros para previsão da energia de falha de empilhamento

Para a previsão da energia de falha de empilhamento, optou-se pelo uso da técnica *Support Vector Regression*. Essa escolha foi baseada no trabalho de Khan et al. [33] no qual os autores utilizaram o SVR para prever a EFE de ligas de alta entropia, alcançando um erro quadrático médio (RMSE) de 24,8 mJ/m². Esses resultados foram combinados com cálculos de endurecimento por solução sólida, aplicando restrições relacionadas a aplicações em altas temperaturas (temperatura *solidus* acima de 1.600 K e intervalo de solidificação inferior a 100 K). Essa abordagem possibilitou a construção de uma fronteira de Pareto para a identificação de ligas promissoras. No contexto de otimização multiobjetivo, a fronteira de Pareto representa o conjunto de soluções com o melhor equilíbrio possível entre duas ou mais propriedades desejadas, nas quais não é possível melhorar um objetivo sem comprometer outro.

Considerando que o presente trabalho utiliza a mesma base de dados de EFE gerada por Khan et al. [33], decidiu-se empregar o mesmo modelo de *machine learning* (SVR). Contudo, diferentemente do estudo original, a técnica de agrupamento k-medoids não foi empregada. Diferentemente do estudo original, a técnica de agrupamento k-medoids não foi aplicada. Além disso, a lista de descritores foi parcialmente reformulada. Enquanto Khan et al. [33] incluíram propriedades como constantes elásticas calculadas por DFT, neste trabalho utilizou-se a lista composta por 36 descritores, conforme detalhado na Seção 4.3 (*Seleção de descritores*) e no APÊNDICE A. Essa adaptação não apenas permitiu alinhar os descritores com os objetivos específicos deste estudo, como também tornou mais eficiente a aplicação da metodologia, uma vez que a mesma lista de descritores já havia sido empregada para a previsão de fase.

Conforme discutido na Seção 3.4.2, os hiperparâmetros mais relevantes para otimização no SVR incluem o parâmetro de regularização *C*, o tipo de *kernel*, *epsilon* e *gamma*. Neste caso, diferentemente do modelo SVM utilizado para previsão de fases, o tempo de processamento não foi considerado uma restrição significativa para a definição das opções ou faixas de cada um desses hiperparâmetros. Devido ao tamanho reduzido da base de dados, composta por apenas 498 pontos, os cálculos puderam ser realizados de forma mais veloz, permitindo a exploração de faixas mais amplas e uma maior diversidade de combinações de hiperparâmetros durante o processo de otimização.

A Tabela 3 apresenta os principais hiperparâmetros do SVR, juntamente com suas respectivas faixas ou opções avaliadas. A seleção dos descritores mais adequados simultaneamente com a otimização dos hiperparâmetros por meio da biblioteca Optuna, seguiu a mesma metodologia aplicada ao modelo SVM, conforme ilustrado esquematicamente na Figura 21.

Tabela 3 - Resumo dos principais hiperparâmetros do modelo SVM e suas respectivas faixas ou opções.

Hiperparâmetro	**Faixa / Opção**
Parâmetro de regularização C	0.001 - 1000
Tipo de Kernel	rbf, poly, sigmoid, linear
Epsilon	0.001 – 1.0
Gama	auto, scale

4.5 Adaptação do algoritmo genético

O algoritmo genético empregado neste trabalho foi adaptado do estudo de Cassar et al. [61], originalmente desenvolvido para o design de composições de cerâmicas vítreas com propriedades específicas: alto índice de refração e baixa temperatura de transição vítrea. Apesar das diferenças nos objetivos entre os dois problemas, ambos compartilham a busca por composições promissoras, característica que exemplifica a flexibilidade dos algoritmos genéticos, conforme destacado na revisão bibliográfica.

As especificidades de cada etapa do algoritmo genético, adaptado para o presente trabalho, serão apresentadas a seguir:

Inicialização da população: A população inicial foi definida como um conjunto de ligas de alta entropia, em que cada indivíduo representa uma liga, seu genoma corresponde à sua composição química e cada gene indica a proporção de um elemento químico. Os elementos considerados foram Al, Co, Cr, Fe, Mn, Ni e V. Para os elementos Co, Cr, Fe, Mn e Ni, as porcentagens poderiam variar entre 0 e 35% atômico, enquanto para Al e V os limites foram definidos entre 0 e 15% atômico. O valor máximo de 35% atômico está em conformidade com a definição de ligas de alta entropia [1,8,9], enquanto o limite de 15% para Al e V reflete a distribuição observada na base de dados de energia de falha de empilhamento (Figura 7), na qual esses elementos não excedem tal percentual.

Assim, a construção da população inicial assegura que os indivíduos explorem adequadamente o espaço composicional relevante.

Avaliação dos indivíduos: Os indivíduos foram avaliados por meio de uma função *fitness*, a qual foi calculada utilizando uma distância euclidiana ponderada, simplificada pela Equação 11. Nessa equação, x e y correspondem ao valor de duas propriedades distintas de um indivíduo, x_d e y_d os valores desejados para essas propriedades, w_x e w_y os pesos que cada propriedade possui para calcular a pontuação do *fitness* e ε_1, ε_2 e ε_3 fatores de penalização caso o requisito da propriedade não seja alcançado.

$$f(x,y) = \sqrt{w_x(x-x_d)^2 + w_y(y-y_d)^2} + \varepsilon_1 + \varepsilon_2 + \varepsilon_3 \qquad (11)$$

Dessa forma, os indivíduos mais aptos apresentam valores menores de *fitness*, indicando que estão mais próximos do objetivo máximo definido.

A função *fitness* foi estruturada considerando múltiplos critérios de avaliação, divididos em duas classes principais:

Predicts: Refere-se a funções cujo objetivo seja maximizar ou minimizar algum parâmetro. Entretanto, faz-se necessário estabelecer limites inferiores e superiores para o valor de cada parâmetro analisado. Isso garante que o algoritmo opere dentro de padrões definidos, limitando o espaço de busca e favorecendo a obtenção de soluções reais e aplicáveis.

Constraints: Consistem em restrições impostas aos parâmetros, nos quais é definido um limite mínimo ou máximo. Contudo, ao contrário dos *predicts*, valores dentro dos critérios estabelecidos não são diferentemente recompensados.

No total, foram considerados cinco critérios na avaliação dos indivíduos.

***Predict* 1- Constante de Hall-Petch (*K*)**

A constante de Hall-Petch está relacionada ao endurecimento por refino de grão e foi calculada por meio da Equação 3. A escolha se deu devido a sua simplicidade e independência de técnicas de processamento específicas, contando com parâmetros que dependem somente do material analisado, permitindo uma aplicação mais ampla.

Embora as simplificações adotadas possam gerar valores que diferem quantitativamente daqueles obtidos experimentalmente, o método é suficiente para identificar qualitativamente as composições com maior potencial para valores elevados de K.

O intervalo estabelecido para maximização do valor de K foi de 100 a 300 MPa.m$^{1/2}$. Os limites foram definidos com base em dados experimentais de ligas reais reportadas na literatura, nas quais os valores de K, quando calculados por meio da equação escolhida, não ultrapassavam 300 MPa·m$^{1/2}$.

***Predict* 2: Tensão de Cisalhamento Crítica Resolvida (τ_Y)**

A tensão de cisalhamento crítica resolvida para endurecimento por solução sólida foi calculada utilizando as Equações 7, 8, 9 e 10. Essa equação é amplamente empregada pelo grupo de pesquisa atual da autora, devido aos seus resultados consistentes para ligas de alta entropia com estrutura CFC [84].

O intervalo definido para maximização do valor de τ_Y foi de 100 a 500 MPa. Essa faixa é consistente com valores reportados na literatura para ligas de alta entropia.

Tanto para o cálculo de K quanto de τ_Y, foram considerados o raio atômico dos elementos e, consequentemente, o seu vetor de Burgers, na estrutura CFC. Estudos como o de Coury et al. [84] demonstraram que resultados mais precisos são obtidos quando as propriedades dos elementos são avaliadas na estrutura final da solução sólida.

Os valores do raio atômico na estrutura CFC foram determinados com base em gráficos que relacionam parâmetro de rede com concentração de soluto, obtidos a partir de estudos de Mishima et al. [85]. Nesse artigo, os autores doparam níquel com outros elementos em concentrações de até 25% atômico, sem alteração da estrutura CFC. A partir dessas curvas, foi possível extrapolar o raio atômico dos dopantes para composições considerando 100% de dopante. Esse método foi aplicado para diversos elementos, incluindo todos os analisados no presente trabalho.

Como exemplo, será descrito o procedimento adotado para a determinação do raio atômico do Mn na estrutura CFC. Considere a Figura 22 obtida do trabalho de Mishima et al. [85], com destaque para a curva do Mn (círculos vermelhos).

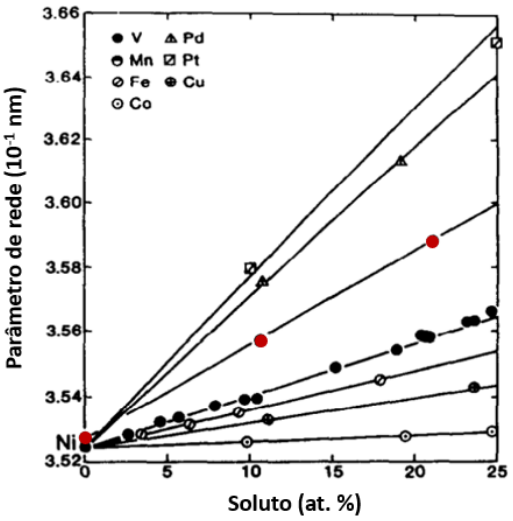

Figura 22 - Variação do parâmetro de rede do níquel (estrutura CFC) em função da adição de diferentes elementos de liga, com destaque para a curva referente ao Mn (círculos vermelhos), utilizada na estimativa do raio atômico por extrapolação para 100% de dopante. Adaptado de Mishima et al. [85].

A curva do Mn pode ser ajusta pela seguinte função linear:

$$y = 0{,}00308289217925x + 3{,}52425101214574$$

Em que y é o parâmetro de rede da liga (em Å) e x é a concentração atômica de Mn. Ao substituir x por 100 (referente à liga hipotética 100% Mn), obtém-se um parâmetro de rede de aproximadamente 3,833 Å, correspondente a um raio atômico de aproximadamente 1,355 Å para o Mn na estrutura CFC. Esse procedimento de extrapolação foi desenvolvido por Gabriela Bugni Ribeiro, até então aluna de graduação em Engenharia de Materiais pela Universidade Federal de São Carlos.

***Predict* 3: Energia de Falha de Empilhamento (EFE)**

A energia de falha de empilhamento foi determinada pelo modelo SVR desenvolvido e otimizado conforme descrito nos subcapítulos 4.2 e 4.4. O objetivo é minimizar o valor da EFE dentro do intervalo de 40 a -100 mJ/m^2. Valores de EFE inferiores a 40 mJ/m² são de particular interesse, pois estão correlacionados com a ativação dos efeitos TWIP/TRIP, os quais promovem o aumento de tenacidade [31].

É importante destacar o debate existente na literatura sobre a obtenção de valores negativos de EFE. Enquanto os valores experimentais disponíveis são sempre positivos, valores negativos têm sido reportados em cálculos baseados na teoria do funcional da densidade (DFT) [86].

Entre os argumentos contrários à existência física de EFE negativas está o entendimento de que a EFE representa um custo energético, e, portanto, deveria ser

sempre positiva. Além disso, uma EFE negativa indicaria que a fase HCP é mais estável que a CFC, cenário no qual o conceito de falha de empilhamento perde sentido, já que a estrutura tenderia à transformação completa para HCP. Alguns autores também apontam que valores negativos podem surgir como artefatos computacionais relacionados à definição do estado de referência [86].

Por outro lado, defensores da validade de EFE negativas sugerem que elas podem ocorrer em sistemas metastáveis e refletir a estabilidade local da fase HCP, sendo influenciadas por fatores como a fricção de rede durante a transformação [86].

Embora o presente trabalho não avalie a existência ou validade de valores negativos, a tendência geral é que, independentemente do sinal, valores mais baixos de EFE indicam uma maior probabilidade de ativação dos efeitos TWIP/TRIP. Dessa forma, a minimização da EFE representa um critério preditivo relevante para o design de ligas de alta entropia.

***Constraint* 1: Previsão de Ligas Monofásicas CFC**

A previsão de ligas monofásicas com estrutura cúbica de face centrada foi realizada utilizando o modelo SVM desenvolvido e otimizado conforme descrito nos subcapítulos 4.2 e 4.4. O objetivo é selecionar ligas classificadas com o número 10, nomenclatura atribuída às ligas monofásicas CFC durante a construção da base de dados.

***Constraint* 2: Porcentagem Atômica dos Elementos**

O último critério se refere à porcentagem atômica dos elementos que compõem as ligas. Os limites estabelecidos foram de 0 a 35% para os elementos Co, Cr, Fe, Mn e Ni, e de 0 a 15% para os elementos Al e V. Esses limites estão fundamentados na definição de ligas de alta entropia [1,8,9], e também na análise da base de dados de EFE utilizada, onde as concentrações máximas de Al e V não ultrapassaram 15% (Figura 7).

Seleção dos indivíduos: A seleção dos indivíduos foi realizada utilizando o método da seleção por torneio, seguindo os parâmetros descritos no trabalho de Cassar et al [61]. Três indivíduos foram selecionados aleatoriamente da população, e aquele com o melhor valor da função *fitness* (ou seja, aquele de menor valor de *fitness*) dentro do grupo foi escolhido para integrar a formulação da próxima geração.

Formação da nova geração: A variabilidade genética entre os indivíduos foi introduzida por meio de cruzamento uniforme e mutação, mantendo os parâmetros descritos no estudo de Cassar et al [61]. No cruzamento uniforme, a probabilidade de cada

gene ser herdado de um dos pais foi fixada em 50%. Para a mutação, cada indivíduo teve uma chance de 20% de sofrer mutação. Caso selecionado, cada gene apresentou uma probabilidade de 5% de alterar seu valor para um número inteiro aleatório dentro do intervalo composicional pré-estabelecido. Esses mecanismos asseguram diversidade genética suficiente para evitar a estagnação prematura do algoritmo.

Finalização do algoritmo genético: O critério de parada do algoritmo foi definido pelo número máximo de gerações. A convergência foi monitorada para assegurar que o processo evolutivo alcançasse soluções otimizadas antes do término das iterações.

Ademais, os próprios parâmetros do algoritmo genético, como o tamanho da população e o número de gerações, foram sistematicamente variados para determinar as configurações ideais capazes de minimizar os valores de *fitness*. Testes foram conduzidos com diferentes combinações de tamanho populacional (100, 200, 500) e número de gerações (100, 200, 500). Cada combinação foi repetida 100 vezes para garantir robustez estatística nos resultados e identificar padrões de convergência confiáveis.

4.6 Active Learning para aprimoramento da previsão da fase

Uma abordagem de *active learning* foi implementada para melhorar a precisão local do modelo SVM utilizado na previsão de fase, conforme necessidade identificada na seção de resultados. O algoritmo genético, adaptado em acorde com o descrito na seção anterior, foi executado 50 vezes, gerando uma lista com 50 ligas candidatas. Cada execução, com tamanho de população e número de gerações igual a 100, selecionava a liga com menor valor de *fitness*.

As ligas selecionadas eram avaliadas quanto à formação de fases por meio de cálculos CALPHAD realizados a 900°C. A integração entre o *script* em Python e o software Pandat™ (base de dados PanHEA2023) foi realizada pelo PanPython, que automatizou a análise das composições geradas pelo algoritmo genético. Essa etapa verificava se as ligas selecionadas pelo algoritmo genético, classificadas como 10 pelo modelo SVM, eram efetivamente monofásicas com estrutura cúbica de face centrada.

Caso pelo menos 90% das ligas fossem confirmadas como monofásicas CFC pelo CALPHAD, o processo era encerrado, indicando que a base de dados do SVM possuía robustez suficiente para atender ao objetivo do trabalho. Se esse critério não fosse alcançado, as 50 composições analisadas eram convertidas em descritores e incorporadas

à base de treinamento do SVM, que passava por um novo treinamento. A seguir, o algoritmo genético era novamente executado para identificar novas ligas candidatas. Esse ciclo iterativo era repetido até que o critério de 90% de acurácia fosse atingido. O fluxograma que representa esse processo pode ser conferido na Figura 23.

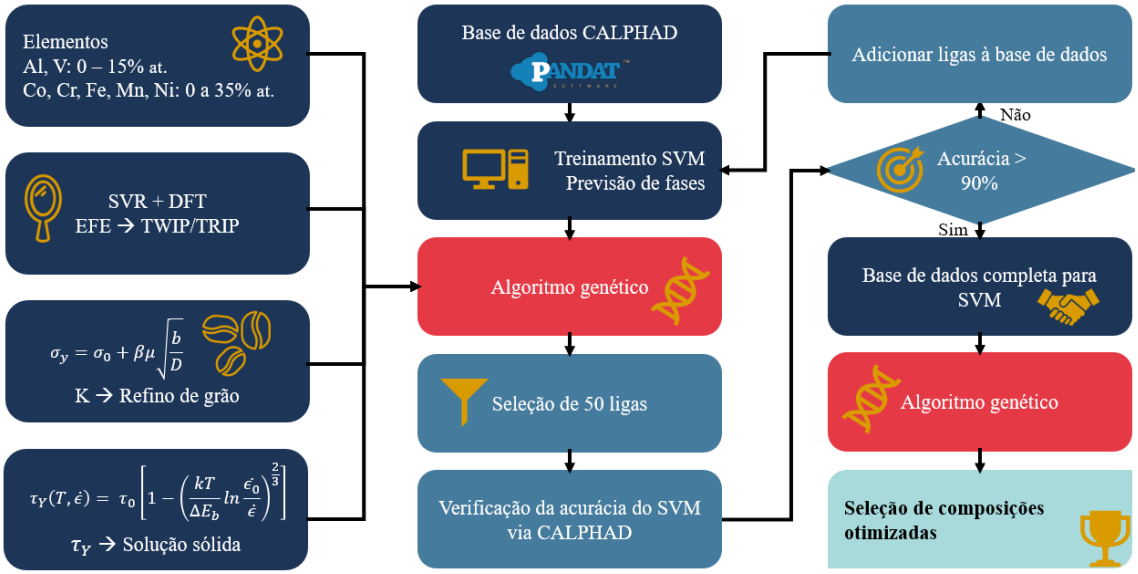

Figura 23 - Fluxograma da implementação do sistema de *active learning* para melhora da acurácia local do modelo de SVM para previsão de ligas monofásicas CFC.

4.7 Seleção de composições de interesse

A versão final do algoritmo genético foi utilizada para identificar duas composições promissoras. Para isso, o algoritmo foi executado 1.000 vezes. Das ligas geradas, foram selecionadas a composição com o menor valor de *fitness* (indicando a configuração mais próxima dos objetivos estabelecidos) e outra com o maior valor de *fitness*

Essa abordagem permitiu maximizar a diferença entre as composições e propriedades previstas, facilitando a análise comparativa. A seleção da composição intermediária também foi essencial para avaliar a eficácia da metodologia em explorar variações no espaço de busca e na identificação de ligas com diferentes combinações de propriedades.

5 RESULTADOS E DISCUSSÃO

5.1 Descritores selecionados e hiperparâmetros otimizados para os modelos de *machine learning*

5.1.1 Descritores selecionados e hiperparâmetros otimizados do modelo SVM para previsão de fase

A seleção sequencial progressiva foi utilizada para identificar o conjunto ideal de descritores para o modelo SVM voltado para previsão de fases. A Figura 24 apresenta a acurácia do modelo em função do número de descritores selecionados. Observa-se que, após a inclusão do oitavo descritor, o valor da acurácia se mantém aproximadamente constante em torno de 97%. Consequentemente, os dez primeiros descritores identificados foram escolhidos como entradas para o treinamento do modelo SVM.

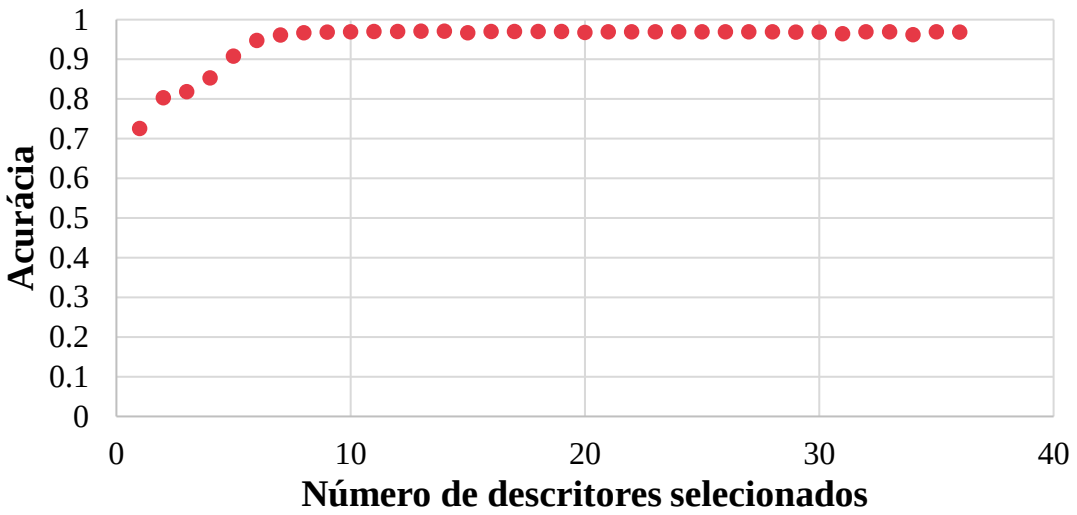

Figura 24 – Acurácia do modelo de SVM para previsão de fase em função do número de descritores selecionados.

Os descritores selecionados, em ordem escolhida, foram:

- Média ponderada da Concentração de Elétrons de Valência;
- Média ponderada do Número de Mendeleev;
- Desvio padrão ponderado da Eletronegatividade;
- Desvio padrão ponderado da Entalpia de Fusão;
- Média ponderada do Módulo de Cisalhamento;
- Média ponderada da Entalpia de Fusão;
- Entropia de Mistura Ideal;

- Desvio padrão ponderado do Período da tabela periódica;
- Soma da Capacidade Térmica de Excesso com a Capacidade Térmica Específica do elemento puro na estrutura CFC a 900 °C;
- Desvio padrão ponderado da Entropia de Fusão.

Os valores otimizados para os hiperparâmetros do modelo foram:

- ***C*** = 916.65;
- ***Kernel*** = rbf;
- ***Gamma*** = scale.

Por se tratar do kernel rbf, não há um valor de *degree* associado.

A acurácia global do modelo com essa configuração foi de 96.98%, demonstrando alta precisão na previsão de fases. A lista completa dos descritores selecionados, juntamente com os valores otimizados de hiperparâmetros, pode ser encontrada no Apêndice A. A matriz de confusão gerada com esses parâmetros é apresentada na Figura 25.

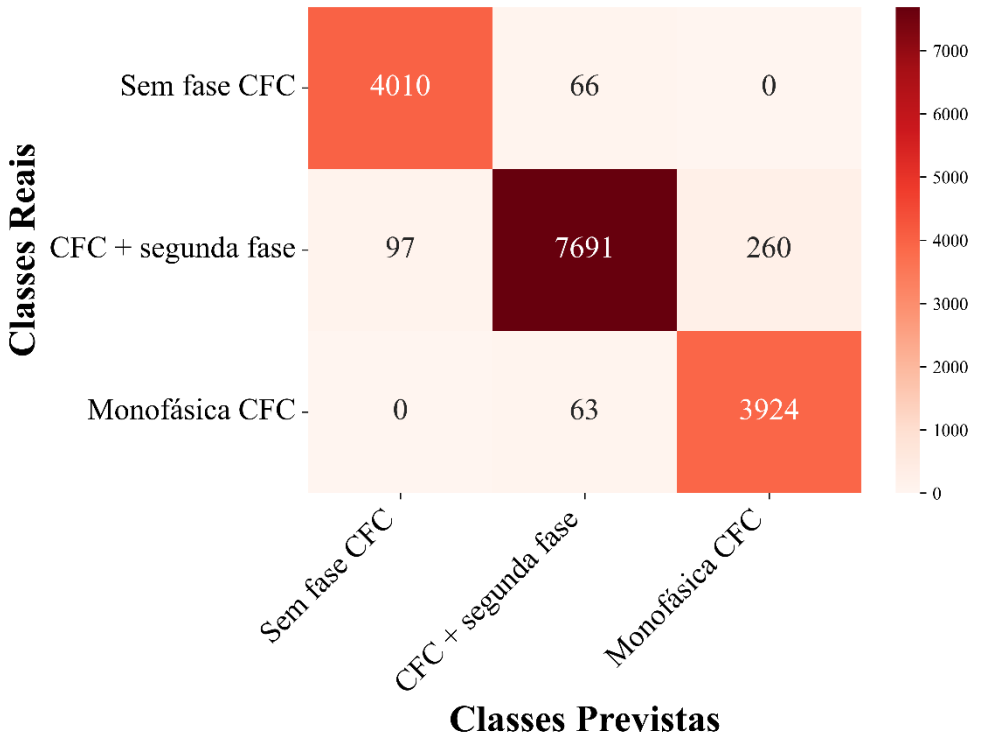

Figura 25 – Matriz de confusão do modelo SVM para previsão de fase após seleção dos dez descritores principais com os hiperparâmetros otimizados.

A análise da matriz de confusão revela que o modelo apresentou excelente desempenho em distinguir as classes. Para a classe de ligas sem fase CFC, o modelo classificou corretamente 4.010 de 4.107 composições (~97,6%), enquanto os erros foram limitados a 66 composições incorretamente atribuídas à classe de ligas CFC + segunda fase. Não houve classificações equivocadas para a classe de ligas monofásicas CFC.

Na classe de ligas CFC + segunda fase, 7.691 de 8.048 composições foram classificadas corretamente (~95,6%). Entretanto, 97 composições foram erroneamente identificadas como pertencentes à classe de ligas sem fase CFC, e 260 composições foram atribuídas incorretamente à classe ligas monofásicas CFC.

Para a classe de ligas monofásicas CFC, 3.924 de 3.987 composições foram corretamente classificadas (~98,4%). Apenas 63 composições foram equivocadamente atribuídas à classe de ligas CFC + segunda fase, enquanto não houve erros para a classe de ligas sem fase CFC.

Esses resultados destacam a robustez do modelo, especialmente na discriminação entre as classes de ligas sem fase CFC e ligas monofásica CFC: percebe-se que nenhuma liga foi falsamente identificada como a outra. Apesar da sobreposição observada entre as classes de ligas CFC + segunda fase e ligas monofásicas CFC, possivelmente atribuída à semelhança nas propriedades físico-químicas dessas classes, a acurácia global do modelo permanece elevada.

5.1.2 Descritores selecionados e hiperparâmetros otimizados do modelo SVR para previsão da energia de falha de empilhamento

A seleção sequencial progressiva foi utilizada para determinar o conjunto ideal de descritores para o modelo SVR voltado para a previsão da energia de falha de empilhamento. A Figura 26 apresenta o erro quadrático médio (RMSE) em função do número de descritores selecionados. Observa-se que, após a inclusão do sexto descritor, o valor de RMSE se estabiliza em aproximadamente 30 mJ/m². Com base nisso, e seguindo o padrão adotado pelo modelo SVM, os dez primeiros descritores identificados foram escolhidos como entradas para o treinamento do modelo SVR.

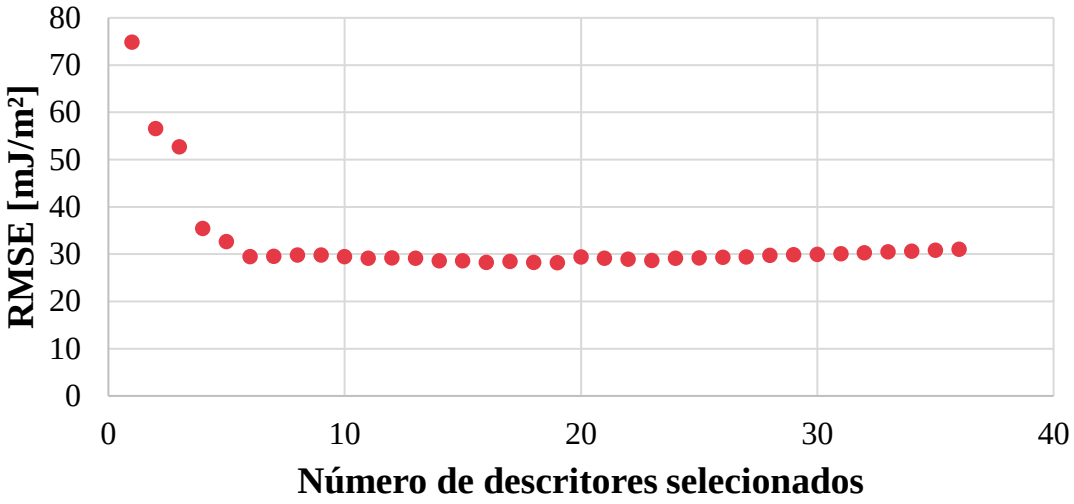

Figura 26 - RMSE do modelo de SVR para previsão da energia de falha de empilhamento em função do número de descritores selecionados.

Os descritores selecionados, em ordem escolhida, foram:

- Entalpia de mistura;
- Desvio padrão ponderado da densidade;
- Média ponderada da entropia de fusão;
- Desvio padrão ponderado da temperatura de ebulição;
- Média ponderada do número de Mendeleev;
- Média ponderada da densidade;
- Média ponderada da eletronegatividade;
- Desvio padrão ponderado da eletronegatividade;
- Desvio padrão ponderado da entropia de fusão;
- Média ponderada do raio atômico.

Os valores otimizados para os hiperparâmetros do modelo foram:

- ***C*** = 926.21;
- ***Kernel*** = rbf;
- ***Epsilon*** = 0.705;
- ***Gamma*** = auto.

O RMSE do modelo com essa configuração foi de 29.45 mJ/m^2. Outros indicadores de desempenho do modelo também foram calculados:

- MAE (Erro Médio Absoluto): 14.59 mJ/m^2 – representa o erro médio absoluto entre os valores previstos e os valores reais;

- MSE (Erro Quadrático Médio): 867.53 mJ^2/m^4 – mede o erro médio ao quadrado, sendo sensível a grandes discrepâncias entre valores reais e previstos;
- R^2 (Coeficiente de Determinação): 0.8829 – indica que aproximadamente 88% da variação nos dados de SFE é explicada pelo modelo.

A lista completa dos descritores selecionados, juntamente com os valores otimizados de hiperparâmetros e os indicadores de desempenho, pode ser encontrada no Apêndice B.

De acordo com a literatura, o efeito TRIP tipicamente ocorre em materiais com valores de EFE inferiores a 20 mJ/m^2. Já o efeito TWIP é observado em materiais com valores intermediários de EFE, entre 20 e 40 mJ/m^2, enquanto valores superiores a 45 mJ/m^2 favorecem predominantemente a deformação plástica por deslizamento de discordâncias [31]. Nesse contexto, o erro RMSE obtido pelo modelo (29.45 mJ/m^2) supera uma faixa completa de classe de EFE, dificultando a precisão na identificação do efeito predominante (TWIP, TRIP ou ambos) nas ligas geradas pelo algoritmo genético.

Entretanto, é importante destacar que os valores de EFE na base de dados utilizada, provenientes de cálculos de DFT, variam amplamente, de aproximadamente -440 mJ/m^2 a 345 mJ/m^2. Assim, apesar de o RMSE parecer significativo em termos absolutos, sua magnitude é pequena em relação à variação total dos dados. Além disso, o algoritmo genético foi desenvolvido com uma função de minimização para EFE, o que aumenta a probabilidade de que as ligas selecionadas apresentem valores baixos de EFE, favorecendo algum mecanismo de endurecimento associado, como TWIP e/ou TRIP.

Dessa forma, a análise dos resultados gerais sugere que o modelo SVR apresenta um equilíbrio eficaz entre simplicidade e precisão, considerando o conjunto de descritores selecionados e os hiperparâmetros otimizados. O valor relativamente baixo de MAE e RMSE, aliado ao elevado R^2 (0,8829), confirma a adequação do modelo para a tarefa de previsão de EFE, especialmente para capturar tendências globais e auxiliar no direcionamento de ligas promissoras.

5.2 Parâmetros intrínsecos do algoritmo genético ajustados

A Figura 27 apresenta gráficos do tipo *boxplot* comparando o impacto do número de gerações e do tamanho da população nos valores de *fitness* do algoritmo genético. Para avaliar o efeito do número de gerações, o tamanho da população foi fixado em 100,

enquanto a análise do impacto do tamanho da população foi realizada mantendo o número de gerações constante em 100.

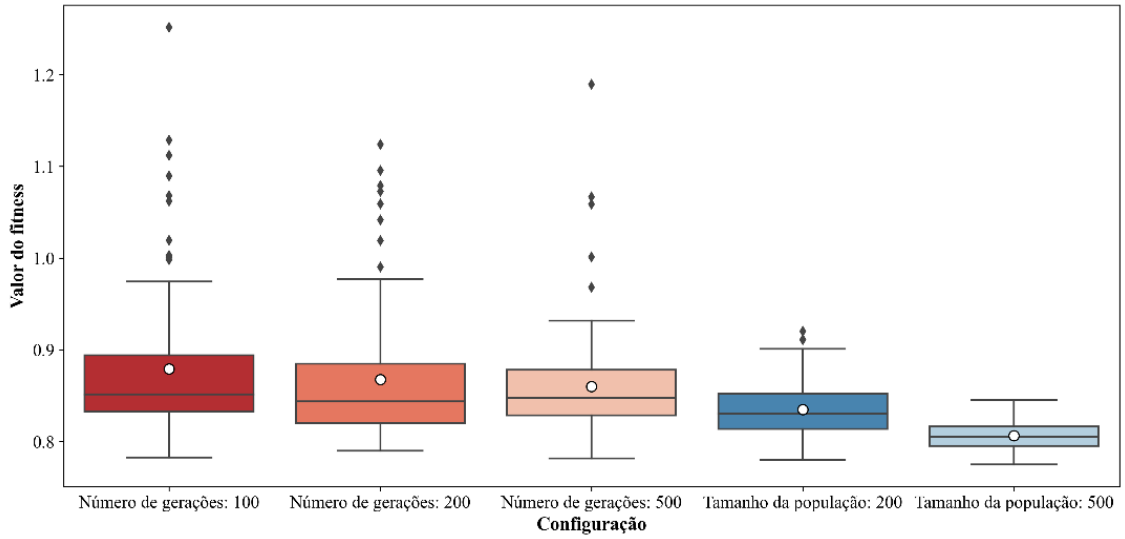

Figura 27 - Boxplot: Influência dos parâmetros do algoritmo genético no valor do *fitness*.

Percebe-se que, à medida que o número de gerações aumenta, o valor médio do *fitness*, representado graficamente pelos círculos brancos localizadas no interior das caixas do gráfico do boxplot, diminui progressivamente de 0,879 para 0,867 e, posteriormente, para 0,860. Em contrapartida, os valores da mediana, representados pela linha contínua no interior das caixas — abaixo da qual se encontram 50% dos dados —, apresentam variações menores e não lineares, passando de 0,851 para 0,844 e, por fim, para 0,847. O intervalo interquartil (IQR), que indica a variabilidade dentro dos 50% centrais dos dados e é representado pela altura de cada caixa, mostra uma tendência de redução conforme o número de gerações aumenta, evidenciando maior consistência dos resultados obtidos.

Quanto ao tamanho da população, é observado que, quanto maior esse valor, menor é o valor médio de *fitness*, decrescendo de 0,835 para 0,806. De forma semelhante, os valores das medianas também diminuem, de 0,831 para 0,805, acompanhados por uma redução no tamanho do IQR. Além disso, houve uma diminuição significativa no número de *outliers* (números atípicos que fogem da média, identificados por losangos pretos) à medida que o tamanho da população aumenta, indicando maior robustez dos resultados gerados.

Essas análises indicam que, para resultados globais mais precisos e consistentes, o aumento do tamanho da população é uma escolha recomendada. No entanto, no presente caso, o objetivo não é obter diversas soluções boas, mas sim identificar uma liga ótima. Apesar das diferenças observadas, todas as configurações parecem convergir para um valor mínimo de *fitness* semelhante, aproximadamente 0,78. Isso sugere que a probabilidade de encontrar um resultado superior para o presente caso está mais associada à variabilidade introduzida pelas diferentes inicializações aleatórias do algoritmo genético do que ao simples aumento do número de gerações ou do tamanho da população.

Consequentemente, para equilibrar o potencial de resultados favoráveis com a eficiência computacional, tanto para a melhoria por *active learning* como na seleção efetiva das composições das ligas de alta entropia, tanto o número de gerações quanto o tamanho da população foram fixados em 100.

5.3 Melhoria promovida pelo método de *active learning*

Após o primeiro teste de simulação, o algoritmo genético identificou 50 ligas promissoras, todas previstas como monofásicas com estrutura CFC a 900°C. Contudo, a verificação subsequente dessas ligas utilizando o software Pandat™ (base de dados PanHEA2023) revelou que apenas 11 das 50 ligas eram realmente monofásicas CFC nas condições especificadas, resultando em uma acurácia de apenas 22%. Este desempenho contrastando fortemente com a acurácia global do modelo SVM, que foi de 96,98%.

Especula-se que a discrepância pode ser atribuída a um aspecto específico das entradas do algoritmo genético, particularmente relacionado ao endurecimento por solução sólida. Para maximizar o limite de escoamento τ_Y, conforme estabelecido nas Equação 7, 8, 9 e 10, é necessário que os raios atômicos dos elementos constituintes da liga apresentem uma grande diferença. Isso aumenta a distorção da rede cristalina, promovendo o aumento do endurecimento por solução sólida. Entretanto, essa diferença significativa no tamanho atômico também representa um obstáculo à formação de uma liga monofásica.

Conforme o algoritmo genético busca identificar ligas com elevados valores de τ_Y que ainda sejam monofásicas CFC, ele tende a explorar uma região do modelo SVM que é mais desafiadora para classificar com precisão.

Para solucionar esse problema, foi empregada uma abordagem de *active learning*. O algoritmo genético foi utilizado para identificar as ligas mais promissoras, destacando pontos de interesse, e consequentemente de difícil classificação, que deveriam ser adicionados à base de dados de treinamento do SVM. Esse processo foi repetido iterativamente, aumentando a representatividade dos dados nessa região do espaço composicional até que a acurácia de previsão para essas condições ultrapassasse 90% (Seção 4.6).

Na primeira iteração do *active learning*, após a inclusão de 50 novas composições na base de dados, a acurácia de previsão de fases aumentou para 46%, com 23 das 50 ligas corretamente classificadas como monofásicas CFC a 900°C. Na segunda iteração, mais 50 dados foram incorporados, resultando em uma melhoria adicional na acurácia, que alcançou 98%, com 49 das 50 ligas corretamente previstas como monofásicas CFC nas mesmas condições.

Ao final do processo iterativo, a base de dados havia sido aprimorada com 100 novas composições, elevando a acurácia local de previsão de fases de 22% para 98%. Esse resultado destaca a eficácia do método de *active learning* na melhoria da precisão do modelo preditivo, especialmente em regiões desafiadoras do espaço composicional.

Além disso, é importante notar que o método não apenas aprimorou a acurácia preditiva, mas também permitiu uma maior compreensão das limitações e potencialidades do modelo SVM. A abordagem demonstrou como o uso iterativo do algoritmo genético para identificar lacunas na base de dados, aliado à incorporação direcionada de novos dados, pode transformar um modelo com limitações específicas em uma ferramenta robusta e confiável para a seleção de ligas.

Esse resultado também reflete o impacto significativo do *active learning* na redução de custos computacionais e experimentais, ao priorizar a aquisição de dados de forma seletiva e eficiente. Essa abordagem é particularmente valiosa em áreas como a ciência de materiais, onde a geração de dados pode ser cara e demorada. Por fim, a implementação bem-sucedida desse método demonstra seu potencial para aplicação em outros problemas de otimização de materiais, como a previsão de outras fases ou propriedades mecânicas.

5.4 Composições selecionadas

Após 1.000 repetições com o número de gerações e o tamanho da população definidos como 100, as composições que apresentaram os valores de *fitness* mais baixo e mais alto foram selecionadas para futura caracterização experimental e estão apresentadas na Tabela 4. A seleção de ligas com valores extremos de *fitness* garante a obtenção de composições distintas, especialmente em relação aos valores EFE, facilitando uma comparação de suas propriedades durante a análise experimental. Os diagramas de fração de fases em função da temperatura para essas ligas estão ilustrados nas Figura 28 e Figura 29.

Tabela 4 – Composições selecionadas via Algoritmo Genético. Os valores de K (constante de Hall-Petch), τ_y (tensão de cisalhamento crítica resolvida), e EFE (energia de falha de empilhamento) são fornecidas em MPa·m$^{1/2}$, MPa e mJ/m^2, respectivamente. As composições são descritas em porcentagem atômica.

Al	Co	Cr	Fe	Mn	Ni	V	fitness	K	τ_y	Phase	EFE
0,98	25,41	11,40	29,97	30,61	1,30	0,33	0,77	241,38	309,85	10	-99,68
1,50	16,21	14,71	18,45	21,45	19,45	8,23	1,77	233,55	303,72	10	32,53

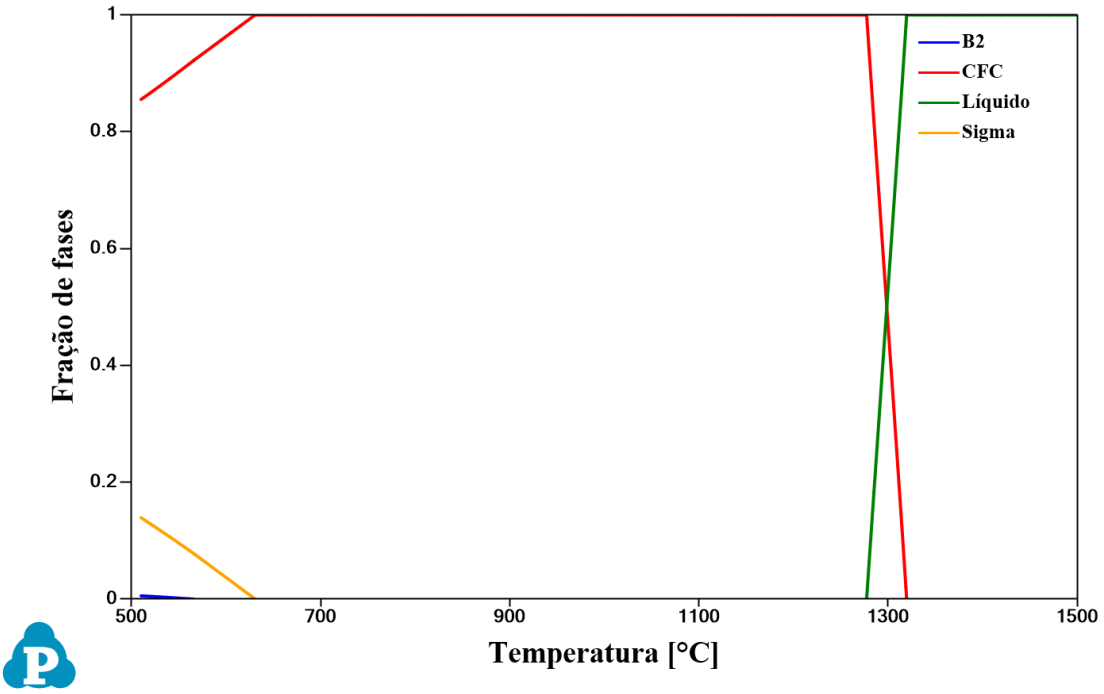

Figura 28 - Diagrama de fração de fases em função da temperatura para a liga de fitness igual a 0,77 obtido via Pandat™ (base de dados PanHEA2023).

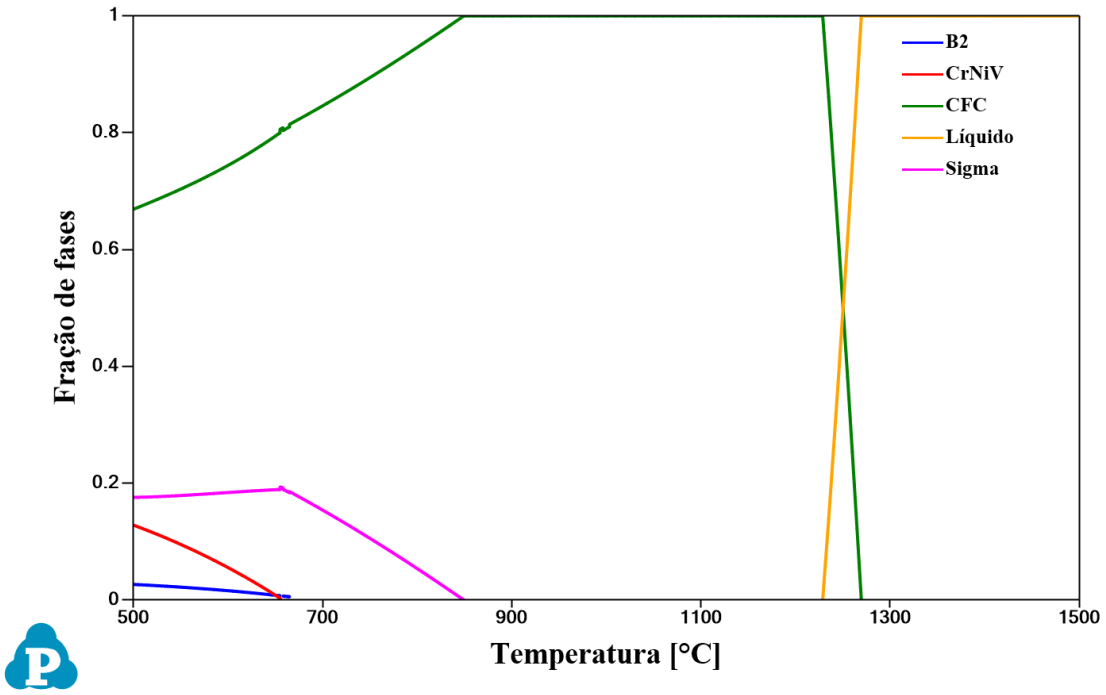

Figura 29 - Diagrama de fração de fases em função da temperatura para a liga de fitness igual a 1,77 obtido via Pandat™ (base de dados PanHEA2023).

A Tabela 5 apresenta as estatísticas gerais das 1.000 composições selecionadas pelo Algoritmo Genético após 1.000 repetições.

Tabela 5 - Estatísticas gerais das 1000 ligas selecionadas pelo Algoritmo Genético após 1000 iterações. Os valores de K (constante de Hall-Petch), τ_y (tensão de cisalhamento crítica resolvida), e EFE (energia de falha de empilhamento) são fornecidas em MPa·m$^{1/2}$, MPa e mJ/m², respectivamente. As composições são descritas em porcentagem atômica. *Média* representa o valor médio da composição ou propriedade, enquanto *Desvio* indica o desvio padrão da respectiva composição ou propriedade.

	Al	Co	Cr	Fe	Mn	Ni	V	fitness	K	τ_y	Phase	EFE
Média	2,75	22,74	13,17	29,81	23,28	7,33	0,91	0,91	238,65	285,95	10,00	-90,54
Desvio	0,83	2,97	3,55	3,34	5,08	4,50	0,79	0,10	3,07	25,83	0,00	14,53

Os dados indicam que a maioria das composições selecionadas apresenta concentrações significativas de Co, Cr, Fe e Mn, com níveis moderados de Ni e menores proporções de Al e V.

Considerando que um dos objetivos primários era minimizar a energia de falha de empilhamento, não é surpreendente que as composições resultantes apresentem elevados teores de Co e baixas porcentagens de Ni. Nesse contexto, Bertoli et al. [14] desenvolveram uma equação empírica para determinar o valor da EFE em ligas Cr-Co-Ni, com base na porcentagem dos elementos presentes. Essa equação permite a

construção de um diagrama ternário que correlaciona as porcentagens dos elementos, o valor da EFE e o mecanismo de acomodação de tensão resultante (Figura 30). Observa-se que maiores porcentagens de Cr e Co levam à redução no valor da energia de falha de empilhamento, enquanto o Ni promove o aumento desse valor. Esses resultados corroboram a eficácia do algoritmo de SVR, indicando que ele está direcionado corretamente na busca por composições com SFE reduzida.

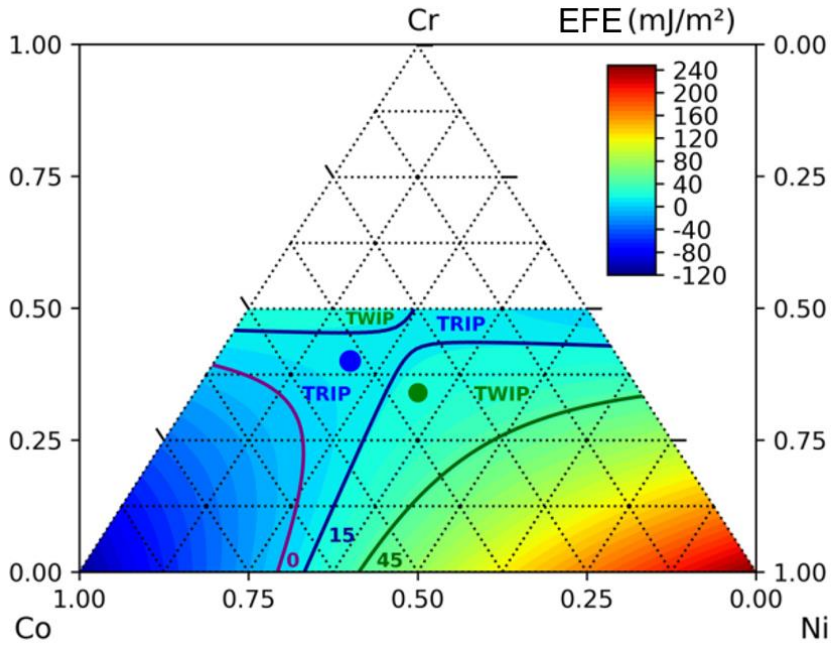

Figura 30 - Energia de falha de empilhamento para ligas Cr-Co-Ni em temperatura ambiente. O ponto em azul representa a liga $Cr_{40}Co_{40}Ni_{20}$ e o ponto verde representa a liga equiatômica CrCoNi [14] (adaptado).

A presença de pequenas quantidades de Al e V, bem abaixo do limite de 15% estabelecido como restrição no Algoritmo Genético, indica que essa restrição não influenciou significativamente os resultados finais. O próprio GA tende a selecionar ligas com porcentagens reduzidas desses elementos.

Os valores resultantes de K (constante de Hall-Petch) mostram variação limitada, com um desvio padrão de 3,07 MPa·$m^{1/2}$, indicando a dificuldade em maximizar esse parâmetro para valores superiores a aproximadamente 240 MPa·$m^{1/2}$. Em contraste, os valores de τ_y (tensão de cisalhamento crítica resolvida) apresentam maior variabilidade, com um desvio padrão de 25,83 MPa, com média de 285,95 MPa.

Quanto aos resultados de fase, o desvio padrão de zero indica que todas as composições identificadas pelo GA foram classificadas como monofásicas CFC a 900°C. Isso reflete a capacidade do algoritmo em cumprir as restrições de fase impostas.

Por fim, os valores de EFE parecem ter sido minimizados tanto quanto possível dentro das restrições definidas. É provável que, se o limite inferior para EFE fosse reduzido, o GA identificaria ligas com valores ainda mais baixos de EFE. Contudo, valores de EFE inferiores a -100 mJ/m² poderiam resultar em ligas cuja estrutura HCP se torna consideravelmente mais estável que a fase CFC, sem a ocorrência do efeito TRIP ou TWIP, o que não é o objetivo deste trabalho.

É importante notar que a segunda liga selecionada difere significativamente dos valores médios (Tabela 5), pois foi escolhida com base no maior valor de *fitness*, tornando-se um *outlier* em relação à maioria das composições selecionadas. Apesar disso, essa liga ainda atende a todas as previsões e restrições definidas na função de *fitness*, evidenciando a robustez do método empregado.

5.5 Importância da integração algoritmo genético com machine learning

A opção por desenvolver uma estrutura de *machine learning* para a identificação de ligas monofásicas CFC, ao invés de integrar diretamente o algoritmo genético ao software termodinâmico Pandat®, foi motivada pela necessidade de otimização computacional e viabilidade prática do projeto. A geração da base de dados bruta para treinamento dos modelos de ML, utilizando cálculos via Pandat®, foi concluída em aproximadamente dois dias. No entanto, a construção do pipeline completo envolveu etapas adicionais como filtragem e preparação dos dados, conversão em descritores, otimização de hiperparâmetros, treinamento, validação cruzada e testes de desempenho dos modelos, o que demandou cerca de três dias ao todo. Ressalta-se que o processo não se limitou a uma execução única: diversas versões da base de dados foram calculadas, múltiplas tentativas de arquitetura de código foram exploradas, e melhorias sucessivas foram implementadas ao longo do tempo. Dessa forma, considerando desde a concepção da metodologia, desenvolvimento dos scripts, testes, até a consolidação do modelo final, o tempo total envolvido pode ser estimado em cerca de seis meses a um ano. Tal período estende-se a até dois anos se considerada a execução completa do mestrado, incluindo revisão bibliográfica e disciplinas obrigatórias.

Em contraste, a integração direta entre o algoritmo genético e o Pandat®, sem o uso de modelos preditivos, implicaria em uma abordagem computacionalmente inviável. A estimativa média de tempo para o cálculo termodinâmico de um único ponto, considerando a criação e manipulação automática dos arquivos necessários para entrada e saída no Pandat®, é de aproximadamente 20 segundos. Dado que o algoritmo genético foi executado com 1000 repetições, cada uma contendo 100 gerações, e 100 ligas por geração, conforme mencionado no tópico 5.4 (composições selecionadas), o número total de cálculos seria da ordem de 10 milhões. Isso resultaria em um tempo computacional estimado em mais de 6 anos contínuos de processamento, mesmo em condições ideais. Portanto, a adoção da metodologia baseada em *machine learning* demonstrou-se não apenas eficiente, mas fundamental para a viabilização prática da proposta deste trabalho.

6 CONCLUSÕES

O objetivo deste trabalho foi desenvolver um algoritmo genético integrado a modelos *machine learning* para o design de ligas de alta entropia com propriedades mecânicas otimizadas. Essa abordagem visa atender às demandas crescentes da ciência de materiais por soluções inovadoras, eficientes e de baixo custo computacional para o desenvolvimento de novas ligas.

Uma das bases fundamentais do trabalho foi o uso do PanPython para a obtenção de bases de dados para previsão de fases. Essa ferramenta demonstrou ser ágil, permitindo a geração rápida de diversas composições por meio de comandos simples. Além disso, a versatilidade do PanPython abre caminho para futuras aplicações, podendo ser adaptado para a obtenção de bases de dados voltadas para a previsão de outras fases ou para o uso de elementos químicos distintos daqueles explorados neste estudo.

O algoritmo de *Support Vector Machine* se revelou uma ferramenta eficiente para a classificação de fases, alcançando uma acurácia global de 96,98%. Esse resultado reforça a confiabilidade do modelo em prever corretamente as fases das ligas estudadas em tempo hábil de execução. Um aspecto fundamental foi a adoção da estratégia de *active learning*, que se mostrou essencial para aumentar a acurácia local do modelo em 98%. Essa abordagem permitiu a identificação de pontos críticos no espaço composicional, otimizando a base de dados de treinamento e reduzindo significativamente os custos computacionais para o aumento da precisão.

Por sua vez, o modelo de *Support Vector Regression*, embora tenha apresentado um erro quadrático médio de 29,45 mJ/m^2 para a previsão da EFE, capturou de forma eficaz as tendências das composições com valores reduzidos deste parâmetro. Esse resultado demonstra a capacidade do SVR em identificar padrões e comportamentos no espaço composicional, mesmo com valores absolutos menos precisos.

Para ambos os modelos de *machine learning*, a transformação das composições brutas em descritores foi um fator determinante. Essa abordagem não apenas aumentou a acurácia dos modelos, mas também possibilita futura extrapolação do código para novos elementos químicos, ampliando o escopo das aplicações. O uso do Optuna foi outro destaque do trabalho, sendo crucial para a seleção de hiperparâmetros otimizados. Essa etapa contribuiu para o aprimoramento da acurácia dos modelos, garantindo que os

algoritmos fossem ajustados de maneira eficiente para atender às especificidades dos dados.

O algoritmo genético, de forma geral, demonstrou ser uma ferramenta poderosa para o design de HEAs. Ele foi capaz de identificar composições promissoras dentro dos limites definidos, resultando em ligas monofásicas CFC, com elevado valor de constante de Hall-Petch e de tensão de cisalhamento crítica resolvida, relacionados com endurecimento por refino de grão e solução sólida, respectivamente, além de valores reduzidos de EFE, visando possível ativação de efeitos TRIP/TWIP. Dessa forma, o algoritmo conseguiu explorar eficientemente um amplo campo composicional e maximizar propriedades frequentemente antagônicas, levando à conclusão de que o projeto atendeu aos objetivos propostos.

7 SUGESTÕES PARA FUTUROS TRABALHOS

Este trabalho abre várias possibilidades de expansão e refinamento para pesquisas futuras no campo do design de ligas de alta entropia, com foco na integração de algoritmos genéticos e *machine learning*. Algumas direções promissoras incluem:

Fabricação das ligas: As ligas selecionadas na Tabela 4 podem ser fabricadas para verificação experimental das propriedades preditas e avaliação da capacidade de otimização do algoritmo genético.

Expansão para Novos Elementos: A abordagem desenvolvida neste trabalho pode ser ampliada para explorar ligas que incluam elementos diferentes dos utilizados aqui, como Cu, Mo e Ti.

Aprimoramento do SVM: A base de dados utilizada no modelo SVR pode ser ampliada para incluir elementos adicionais como Cu, Mo e Ti. Para tanto, poderia ser utilizada a interface do PanPython com outras bases de dados que não a PanHEA2023. Poderiam ser ainda consideradas outras faixas de temperaturas ou fases para a análise.

Aprimoramento do SVR: A base de dados utilizada no modelo SVR pode ser ampliada com novos cálculos de DFT, tanto para incluir elementos adicionais, como Cu, Mo e Ti, quanto para criar sistemas de ordem variável, explorando ligas com diferentes números de elementos simultâneos (diferentes de sete). Isso permitiria capturar tendências em sistemas mais complexos e potencialmente aumentar a precisão do modelo.

Cálculo Alternativo de Efeitos TRIP/TWIP: A tendência à ativação dos efeitos TRIP/TWIP pode ser investigada utilizando métodos alternativos, como a diferença na energia livre de Gibbs entre as fases CFC e HCP. Essa abordagem poderia fornecer *insights* adicionais sobre o comportamento mecânico das ligas e sua relação com a energia de falha de empilhamento.

Refinamento de Fórmulas para K e τ_y: Novas fórmulas mais complexas, incorporando parâmetros adicionais, poderiam ser testadas para a definição dos valores da constante de Hall-Petch e da tensão de cisalhamento crítica resolvida. Isso permitiria uma descrição mais detalhada e precisa das propriedades mecânicas das ligas projetadas.

Exploração de Parâmetros do Algoritmo Genético: Diferentes taxas de cruzamento e mutação podem ser investigadas no algoritmo genético, permitindo um melhor entendimento de como esses parâmetros influenciam a eficiência e a eficácia na exploração do espaço composicional. Além disso, o próprio algoritmo pode ser adaptado

para incluir novas restrições, como: custo, facilidade de fabricação ou conformação das ligas, impactos ambientais, dentre outros, potencialmente promovendo resultados mais aplicáveis industrialmente.

8 REFERÊNCIAS BIBLIOGRÁFICAS

[1] G. Bertoli, V.G.L. De Sousa, D. De A. Santana, L.B. Otani, C.S. Kiminami, F.G. Coury, Phase equilibria of VCrMnFeCo high entropy alloys, Journal of Alloys and Compounds 903 (2022) 163950. https://doi.org/10.1016/j.jallcom.2022.163950.

[2] Y.-F. Yang, F. Hu, T. Xia, R.-H. Li, J.-Y. Bai, J.-Q. Zhu, J.-Y. Xu, G.-F. Zhang, High entropy alloys: A review of preparation techniques, properties and industry applications, Journal of Alloys and Compounds 1010 (2025) 177691. https://doi.org/10.1016/j.jallcom.2024.177691.

[3] B. Gludovatz, A. Hohenwarter, K.V.S. Thurston, H. Bei, Z. Wu, E.P. George, R.O. Ritchie, Exceptional damage-tolerance of a medium-entropy alloy CrCoNi at cryogenic temperatures, Nat Commun 7 (2016) 10602. https://doi.org/10.1038/ncomms10602.

[4] G. Renner, A. Ekárt, Genetic algorithms in computer aided design, Computer-Aided Design 35 (2003) 709–726. https://doi.org/10.1016/S0010-4485(03)00003-4.

[5] Y.F. Ye, Q. Wang, J. Lu, C.T. Liu, Y. Yang, Design of high entropy alloys: A single-parameter thermodynamic rule, Scripta Materialia 104 (2015) 53–55. https://doi.org/10.1016/j.scriptamat.2015.03.023.

[6] S. Guo, C. Ng, J. Lu, C.T. Liu, Effect of valence electron concentration on stability of fcc or bcc phase in high entropy alloys, Journal of Applied Physics 109 (2011). https://doi.org/10.1063/1.3587228.

[7] G. Cacciamani, AN INTRODUCTION TO THE CALPHAD METHOD AND THE COMPOUND ENERGY FORMALISM (CEF), TMM 13 (2016) 16–24. https://doi.org/10.4322/2176-1523.1048.

[8] F.G. Coury, G. Zepon, C. Bolfarini, Multi-principal element alloys from the CrCoNi family: outlook and perspectives, Journal of Materials Research and Technology 15 (2021) 3461–3480. https://doi.org/10.1016/j.jmrt.2021.09.095.

[9] D.B. Miracle, High entropy alloys as a bold step forward in alloy development, Nat Commun 10 (2019) 1805. https://doi.org/10.1038/s41467-019-09700-1.

[10] B. Cantor, I.T.H. Chang, P. Knight, A.J.B. Vincent, Microstructural development in equiatomic multicomponent alloys, Materials Science and Engineering: A 375–377 (2004) 213–218. https://doi.org/10.1016/j.msea.2003.10.257.

[11] J. -W. Yeh, S. -K. Chen, S. -J. Lin, J. -Y. Gan, T. -S. Chin, T. -T. Shun, C. -H. Tsau, S. -Y. Chang, Nanostructured High-Entropy Alloys with Multiple Principal Elements: Novel Alloy Design Concepts and Outcomes, Adv Eng Mater 6 (2004) 299–303. https://doi.org/10.1002/adem.200300567.

[12] D.B. Miracle, O.N. Senkov, A critical review of high entropy alloys and related concepts, Acta Materialia 122 (2017) 448–511. https://doi.org/10.1016/j.actamat.2016.08.081.

[13] J.-W. Yeh, Physical Metallurgy of High-Entropy Alloys, JOM 67 (2015) 2254–2261. https://doi.org/10.1007/s11837-015-1583-5.

[14] G. Bertoli, L.B. Otani, A.J. Clarke, C.S. Kiminami, F.G. Coury, Hall–Petch and grain growth kinetics of the low stacking fault energy TRIP Cr40Co40Ni20 multi-principal element alloy, Applied Physics Letters 119 (2021) 061903. https://doi.org/10.1063/5.0057888.

[15] E.O. Hall, The Deformation and Ageing of Mild Steel: III Discussion of Results, Proc. Phys. Soc. B 64 (1951) 747–753. https://doi.org/10.1088/0370-1301/64/9/303.

[16] N.J. Petch, The Cleavage Strength of Polycrystals, Journal of the Iron and Steel Institute 174 (1953) 25–28.

[17] Z.C. Cordero, B.E. Knight, C.A. Schuh, Six decades of the Hall–Petch effect – a survey of grain-size strengthening studies on pure metals, International Materials Reviews 61 (2016) 495–512. https://doi.org/10.1080/09506608.2016.1191808.

[18] R.B. Figueiredo, M. Kawasaki, T.G. Langdon, Seventy years of Hall-Petch, ninety years of superplasticity and a generalized approach to the effect of grain size on flow stress, Progress in Materials Science 137 (2023) 101131. https://doi.org/10.1016/j.pmatsci.2023.101131.

[19] J.C.M. Li, Y.T. Chou, The role of dislocations in the flow stress grain size relationships, Metall Trans 1 (1970) 1145–1159. https://doi.org/10.1007/BF02900225.

[20] M.F. Ashby, The deformation of plastically non-homogeneous materials, The Philosophical Magazine: A Journal of Theoretical Experimental and Applied Physics 21 (1970) 399–424. https://doi.org/10.1080/14786437008238426.

[21] B.E. M'Rabat, L. Priester, Influence of the Grain Boundary Chemistry on the Extrinsic Dislocation Accommodation in Iron, Materials Science and Engineering: A 125 (1990) 31–38. https://doi.org/doi.org/10.1016/0921-5093(90)90249-3.

[22] W.D. Callister, D.G. Rethwisch, Materials Science and Engineering, 8th ed., Wiley, 2010.

[23] G.E. Dieter, Mechanical Metallurgy, 1988.

[24] W. Hume-Rothery, The structure of metals and alloys, Metals & Metallurgy Trust, 1969. https://archive.org/details/structureofmetal0000hume/page/n443/mode/2up.

[25] C. Varvenne, A. Luque, W.A. Curtin, Theory of strengthening in fcc high entropy alloys, Acta Materialia 118 (2016) 164–176. https://doi.org/10.1016/j.actamat.2016.07.040.

[26] C. Varvenne, G.P.M. Leyson, M. Ghazisaeidi, W.A. Curtin, Solute strengthening in random alloys, Acta Materialia 124 (2017) 660–683. https://doi.org/10.1016/j.actamat.2016.09.046.

[27] ASTM International, ASTM E8/E8M - Standard Test Methods for Tension Testing of Metallic Materials, (n.d.). https://doi.org/10.1520/E0008_E0008M-13A.

[28] D. Hull, D.J. Bacon, eds., Introduction to dislocations, 5th ed, Butterworth-Heinemann, Oxford, 2011.

[29] R.S. Mishra, R.S. Haridas, P. Agrawal, High entropy alloys – Tunability of deformation mechanisms through integration of compositional and microstructural domains, Materials Science and Engineering: A 812 (2021) 141085. https://doi.org/10.1016/j.msea.2021.141085.

[30] R.E. Reed-Hill, Princípios de Metalurgia Física, 2nd ed., Guanabara Dois, 1982.

[31] S.L. Wong, M. Madivala, U. Prahl, F. Roters, D. Raabe, A crystal plasticity model for twinning- and transformation-induced plasticity, Acta Materialia 118 (2016) 140–151. https://doi.org/10.1016/j.actamat.2016.07.032.

[32] T.-H. Lee, E. Shin, C.-S. Oh, H.-Y. Ha, S.-J. Kim, Correlation between stacking fault energy and deformation microstructure in high-interstitial-alloyed austenitic steels, Acta Materialia 58 (2010) 3173–3186. https://doi.org/10.1016/j.actamat.2010.01.056.

[33] T.Z. Khan, T. Kirk, G. Vazquez, P. Singh, A.V. Smirnov, D.D. Johnson, K. Youssef, R. Arróyave, Towards stacking fault energy engineering in FCC high entropy alloys, Acta Materialia 224 (2022) 117472. https://doi.org/10.1016/j.actamat.2021.117472.

[34] D.S. Sholl, J.A. Steckel, Density Functional Theory: A Practical Introduction, 1st ed., Wiley, 2009. https://doi.org/10.1002/9780470447710.

[35] M.E. Tuckerman, Statistical mechanics: theory and molecular simulation, Oxford University Press, Oxford ; New York, 2010.

[36] U.R. Kattner, THE CALPHAD METHOD AND ITS ROLE IN MATERIAL AND PROCESS DEVELOPMENT, TMM 13 (2016) 3–15. https://doi.org/10.4322/2176-1523.1059.

[37] A.A. Deshmukh, R. Ranganathan, Recent advances in modelling structure-property correlations in high-entropy alloys, Journal of Materials Science & Technology 204 (2025) 127–151. https://doi.org/10.1016/j.jmst.2024.03.027.

[38] Z.-K. Liu, Computational thermodynamics and its applications, Acta Materialia 200 (2020) 745–792. https://doi.org/10.1016/j.actamat.2020.08.008.

[39] P.I. Odetola, B.J. Babalola, A.E. Afolabi, U.S. Anamu, E. Olorundaisi, M.C. Umba, T. Phahlane, O.O. Ayodele, P.A. Olubambi, Exploring high entropy alloys: A review on thermodynamic design and computational modeling strategies for advanced materials applications, Heliyon 10 (2024) e39660. https://doi.org/10.1016/j.heliyon.2024.e39660.

[40] C. Zhang, Y. Yang, The CALPHAD approach for HEAs: Challenges and opportunities, MRS Bulletin 47 (2022) 158–167. https://doi.org/10.1557/s43577-022-00284-8.

[41] O.N. Senkov, J.D. Miller, D.B. Miracle, C. Woodward, Accelerated exploration of multi-principal element alloys with solid solution phases, Nat Commun 6 (2015) 6529. https://doi.org/10.1038/ncomms7529.

[42] O.N. Senkov, J.D. Miller, D.B. Miracle, C. Woodward, Accelerated exploration of multi-principal element alloys for structural applications, Calphad 50 (2015) 32–48. https://doi.org/10.1016/j.calphad.2015.04.009.

[43] C. Janiesch, P. Zschech, K. Heinrich, Machine learning and deep learning, Electron Markets 31 (2021) 685–695. https://doi.org/10.1007/s12525-021-00475-2.

[44] J. Wei, X. Chu, X. Sun, K. Xu, H. Deng, J. Chen, Z. Wei, M. Lei, Machine learning in materials science, InfoMat 1 (2019) 338–358. https://doi.org/10.1002/inf2.12028.

[45] J. He, Z. Li, P. Zhao, H. Zhang, F. Zhang, L. Wang, X. Cheng, Machine learning-assisted design of high-entropy alloys with superior mechanical properties, Journal of

Materials Research and Technology 33 (2024) 260–286. https://doi.org/10.1016/j.jmrt.2024.09.014.

[46] S. Elkatatny, W. Abd-Elaziem, T.A. Sebaey, M.A. Darwish, A. Hamada, Machine-learning synergy in high-entropy alloys: A review, Journal of Materials Research and Technology 33 (2024) 3976–3997. https://doi.org/10.1016/j.jmrt.2024.10.034.

[47] M. Awad, R. Khanna, Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers, 2015.

[48] W. Ben Chaabene, M. Flah, M.L. Nehdi, Machine learning prediction of mechanical properties of concrete: Critical review, Construction and Building Materials 260 (2020) 119889. https://doi.org/10.1016/j.conbuildmat.2020.119889.

[49] D. Vishwakarma, V.S.N. Neigapula, Prediction of phase via machine learning in high entropy alloys, Materials Today: Proceedings (2023) S2214785323026871. https://doi.org/10.1016/j.matpr.2023.05.065.

[50] A. Bansal, P. Kumar, S. Yadav, V.S. Hariharan, R. M R, G. Phanikumar, Accelerated design of high entropy alloys by integrating high throughput calculation and machine learning, Journal of Alloys and Compounds 960 (2023) 170543. https://doi.org/10.1016/j.jallcom.2023.170543.

[51] S. Swateelagna, M. Singh, M.R. Rahul, Explainable Machine Learning based approach for the design of new refractory high entropy alloys, Intermetallics 167 (2024) 108198. https://doi.org/10.1016/j.intermet.2024.108198.

[52] N. Radhika, M.S. Niketh, U.V. Akhil, A.A. Adediran, T.-C. Jen, High entropy alloys for hydrogen storage applications: A machine learning-based approach, Results in Engineering 23 (2024) 102780. https://doi.org/10.1016/j.rineng.2024.102780.

[53] A.A. Catal, E. Bedir, R. Yilmaz, M.A. Swider, C. Lee, O. El-Atwani, H.J. Maier, H.C. Ozdemir, D. Canadinc, Machine learning assisted design of novel refractory high entropy alloys with enhanced mechanical properties, Computational Materials Science 231 (2024) 112612. https://doi.org/10.1016/j.commatsci.2023.112612.

[54] S. Jain, R. Jain, V. Kumar, S. Samal, Data-driven design of high bulk modulus high entropy alloys using machine learning, Journal of Alloys and Metallurgical Systems 8 (2024) 100128. https://doi.org/10.1016/j.jalmes.2024.100128.

[55] Y. Tian, D. Xue, R. Yuan, Y. Zhou, X. Ding, J. Sun, T. Lookman, Efficient estimation of material property curves and surfaces via active learning, Phys. Rev. Materials 5 (2021) 013802. https://doi.org/10.1103/PhysRevMaterials.5.013802.

[56] A. Jose, E. Devijver, N. Jakse, R. Poloni, Informative Training Data for Efficient Property Prediction in Metal–Organic Frameworks by Active Learning, J. Am. Chem. Soc. 146 (2024) 6134–6144. https://doi.org/10.1021/jacs.3c13687.

[57] G.A. Sulley, J. Raush, M.M. Montemore, J. Hamm, Accelerating high-entropy alloy discovery: efficient exploration via active learning, Scripta Materialia 249 (2024) 116180. https://doi.org/10.1016/j.scriptamat.2024.116180.

[58] A. Bondu, V. Lemaire, M. Boulle, Exploration vs. exploitation in active learning : A Bayesian approach, in: The 2010 International Joint Conference on Neural Networks (IJCNN), IEEE, Barcelona, Spain, 2010: pp. 1–7. https://doi.org/10.1109/IJCNN.2010.5596815.

[59] S. Katoch, S.S. Chauhan, V. Kumar, A review on genetic algorithm: past, present, and future, Multimed Tools Appl 80 (2021) 8091–8126. https://doi.org/10.1007/s11042-020-10139-6.

[60] O. Velez-Langs, Genetic algorithms in oil industry: An overview, Journal of Petroleum Science and Engineering 47 (2005) 15–22. https://doi.org/10.1016/j.petrol.2004.11.006.

[61] D.R. Cassar, G.G. Santos, E.D. Zanotto, Designing optical glasses by machine learning coupled with a genetic algorithm, Ceramics International 47 (2021) 10555–10564. https://doi.org/10.1016/j.ceramint.2020.12.167.

[62] S. Mirjalili, Genetic Algorithm, in: Evolutionary Algorithms and Neural Networks, Springer International Publishing, Cham, 2019: pp. 43–55. https://doi.org/10.1007/978-3-319-93025-1_4.

[63] M. Kumar, M. Husain, N. Upreti, D. Gupta, Genetic Algorithm: Review and Application, SSRN Journal (2010). https://doi.org/10.2139/ssrn.3529843.

[64] A. Lambora, K. Gupta, K. Chopra, Genetic Algorithm- A Literature Review, in: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), IEEE, Faridabad, India, 2019: pp. 380–384. https://doi.org/10.1109/COMITCon.2019.8862255.

[65] E. Wirsansky, Hands-on genetic algorithms with Python: applying genetic algorithms to solve real-world deep learning and artificial intelligence problems, 1st ed, Packt Publishing, Erscheinungsort nicht ermittelbar, 2020.

[66] S.N. Sivanandam, S.N. Deepa, eds., Introduction to genetic algorithms, Springer, Berlin Heidelberg, 2008.

[67] E. Menou, I. Toda-Caraballo, P.E.J. Rivera-Díaz-del-Castillo, C. Pineau, E. Bertrand, G. Ramstein, F. Tancret, Evolutionary design of strong and stable high entropy alloys using multi-objective optimisation based on physical models, statistics and thermodynamics, Materials & Design 143 (2018) 185–195. https://doi.org/10.1016/j.matdes.2018.01.045.

[68] Y. Ikeda, A New Method of Alloy Design Using a Genetic Algorithm and Molecular Dynamics Simulations and Its Application to Nickel-Based Superalloys, Materials Transactions 38 (1997) 771–779.

[69] G.H. Jóhannesson, T. Bligaard, A.V. Ruban, H.L. Skriver, K.W. Jacobsen, J.K. Nørskov, Combined Electronic Structure and Evolutionary Search Approach to Materials Design, Phys. Rev. Lett. 88 (2002) 255506. https://doi.org/10.1103/PhysRevLett.88.255506.

[70] PanPhaseDiagram, CompuTherm (2020). https://computherm.com/panphasediagram (accessed January 9, 2025).

[71] Pandat SDKs, CompuTherm (2021). https://computherm.com/sdk (accessed January 9, 2025).

[72] Y. Zeng, M. Man, K. Bai, Y.-W. Zhang, Revealing high-fidelity phase selection rules for high entropy alloys: A combined CALPHAD and machine learning study, Materials & Design 202 (2021) 109532. https://doi.org/10.1016/j.matdes.2021.109532.

[73] I. Tanaka, ed., Nanoinformatics, Springer Singapore, Singapore, 2018. https://doi.org/10.1007/978-981-10-7617-6.

[74] Y. Zhang, Y.J. Zhou, J.P. Lin, G.L. Chen, P.K. Liaw, Solid-Solution Phase Formation Rules for Multi-component Alloys, Adv Eng Mater 10 (2008) 534–538. https://doi.org/10.1002/adem.200700240.

[75] Z. Zhou, Y. Zhou, Q. He, Z. Ding, F. Li, Y. Yang, Machine learning guided appraisal and exploration of phase design for high entropy alloys, Npj Comput Mater 5 (2019) 128. https://doi.org/10.1038/s41524-019-0265-1.

[76] G. Deffrennes, B. Hallstedt, T. Abe, Q. Bizot, E. Fischer, J.-M. Joubert, K. Terayama, R. Tamura, Data-driven study of the enthalpy of mixing in the liquid phase, Calphad 87 (2024) 102745. https://doi.org/10.1016/j.calphad.2024.102745.

[77] The Photographic Periodic Table of the Elements, (n.d.). https://periodictable.com/ (accessed January 11, 2025).

[78] L. Ward, A general-purpose machine learning framework for predicting, Npj Computational Materials (2016).

[79] D.W. Aha, R.L. Bankert, A Comparative Evaluation of Sequential Feature Selection Algorithms, in: D. Fisher, H.-J. Lenz (Eds.), Learning from Data, Springer New York, New York, NY, 1996: pp. 199–206. https://doi.org/10.1007/978-1-4612-2404-4_19.

[80] K.R. Santos, Classificação Preditiva de Fases Para Ligas Multicomponentes CrCoFeMnNi Utilizando Machine Learning, Trabalho de Conclusão de Curso, Universidade Federal de São Carlos, 2023.

[81] SVC, Scikit-Learn (n.d.). https://scikit-learn/stable/modules/generated/sklearn.svm.SVC.html (accessed January 12, 2025).

[82] A. Gholamy, V. Kreinovich, O. Kosheleva, Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation, (n.d.).

[83] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A Next-generation Hyperparameter Optimization Framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, Anchorage AK USA, 2019: pp. 2623–2631. https://doi.org/10.1145/3292500.3330701.

[84] F.G. Coury, K.D. Clarke, C.S. Kiminami, M.J. Kaufman, A.J. Clarke, High Throughput Discovery and Design of Strong Multicomponent Metallic Solid Solutions, Sci Rep 8 (2018) 8600. https://doi.org/10.1038/s41598-018-26830-6.

[85] Y. Mishima, S. Oshiai, T. Suzuki, Lattice parameters of Ni(γ), Ni3Al(γ') and Ni3Ga(γ') solid solutions with additions of transition and B-subgroup elements, Acta Metallurgica 33 (1985) 1161–1169. https://doi.org/10.1016/0001-6160(85)90211-1.

[86] X. Sun, S. Lu, R. Xie, X. An, W. Li, T. Zhang, C. Liang, X. Ding, Y. Wang, H. Zhang, L. Vitos, Can experiment determine the stacking fault energy of metastable alloys?, Materials & Design 199 (2021) 109396. https://doi.org/10.1016/j.matdes.2020.109396.

APÊNDICE A

Tabela A.1 – Equações dos descritores empregados na engenharia de características das bases de dados para previsão de fases e energia de falha de empilhamento.

Nome do descritor	Fórmula	Explicação
VEC_avgs_w Média ponderada da concentração de elétrons de valência	$VEC_{avgs\_w} = \sum VEC_i \cdot \omega_i$	VEC_i: Concentração de elétrons de valência do elemento *i* ω_i: Fração molar do elemento *i* na composição
VEC_std_devs_w Desvio padrão ponderado da concentração de elétrons de valência	$VEC_{std\_devs\_w} = \sqrt{\sum (VEC_i - \overline{VEC})^2 \cdot \omega_i}$	VEC_i: Concentração de elétrons de valência do elemento *i* $\overline{VEC}$*:* Média ponderada da concentração de elétrons de valência ω_i: Fração molar do elemento *i* na composição
Density_avgs_w Média ponderada da densidade	$\rho_{avgs\_w} = \sum \rho_i \cdot \omega_i$	ρ_i: Densidade do elemento *i* ω_i: Fração molar do elemento *i* na composição
Density_std_devs_w Desvio padrão ponderado da densidade	$\rho_{std\_devs\_w} = \sqrt{\sum (\rho_i - \bar{\rho})^2 \cdot \omega_i}$	ρ_i: Densidade do elemento *i* $\bar{\rho}$*:* Média ponderada da densidade ω_i: Fração molar do elemento *i* na composição
Electronegativity_avgs_w Média ponderada da eletronegatividade	$\chi_{avgs\_w} = \sum \chi_i \cdot \omega_i$	χ_i: Eletronegatividade do elemento *i* ω_i: Fração molar do elemento *i* na composição
Electronegativity _std_devs_w Desvio padrão ponderado da eletronegatividade	$\chi_{std\_devs\_w} = \sqrt{\sum (\chi_i - \bar{\chi})^2 \cdot \omega_i}$	χ_i: Eletronegatividade do elemento *i* $\bar{\chi}$*:* Média ponderada da eletronegatividade ω_i: Fração molar do elemento *i* na composição
Hfus_avgs_w Média ponderada da entalpia de fusão	$H_{fus\_avgs\_w} = \sum H_{fus,i} \cdot \omega_i$	$H_{fus,i}$: Entalpia de fusão do elemento *i*

		ω_i: Fração molar do elemento *i* na composição
Hfus_std_devs_w Desvio padrão ponderado da entalpia de fusão	$H_{fus\_std\_devs\_w} = \sqrt{\sum \left(H_{fus,i} - \overline{H_{fus}}\right)^2 \cdot \omega_i}$	$H_{fus,i}$: Entalpia de fusão do elemento *i* $\overline{H_{fus}}$: Média ponderada da entalpia de fusão ω_i: Fração molar do elemento *i* na composição
Sfus_avgs_w Média ponderada da entropia de fusão	$S_{fus\_avgs\_w} = \sum S_{fus,i} \cdot \omega_i$	$S_{fus,i}$: Entropia de fusão do elemento *i* ω_i: Fração molar do elemento *i* na composição
Sfus_std_devs_w Desvio padrão ponderado da entropia de fusão	$S_{fus\_std\_devs\_w} = \sqrt{\sum \left(S_{fus,i} - \overline{S_{fus}}\right)^2 \cdot \omega_i}$	$S_{fus,i}$: Entropia de fusão do elemento *i* $\overline{S_{fus}}$: Média ponderada da entropia de fusão ω_i: Fração molar do elemento *i* na composição
Column_avgs_w Média ponderada do grupo da tabela periódica	$C_{avgs\_w} = \sum C_i \cdot \omega_i$	C_i: Grupo da tabela periódica do elemento *i* ω_i: Fração molar do elemento *i* na composição
Column _std_devs_w Desvio padrão ponderado do grupo da tabela periódica	$C_{std\_devs\_w} = \sqrt{\sum (C_i - \bar{C})^2 \cdot \omega_i}$	C_i: Grupo da tabela periódica do elemento *i* $\bar{C}$: Média ponderada do grupo da tabela periódica ω_i: Fração molar do elemento *i* na composição
AtomicWeight_avgs_w Média ponderada da massa atômica	$MM_{avgs\_w} = \sum MM_i \cdot \omega_i$	MM_i: Massa atômica do elemento *i* ω_i: Fração molar do elemento *i* na composição
AtomicWeight _std_devs_w	$MM_{std\_devs\_w} = \sqrt{\sum (MM_i - \overline{MM})^2 \cdot \omega_i}$	MM_i: Massa atômica do elemento *i* $\overline{MM}$: Média ponderada da massa atômica

Desvio padrão ponderado da massa atômica		ω_i: Fração molar do elemento *i* na composição
Shear_avgs_w Média ponderada do módulo de cisalhamento	$\mu_{avgs\_w} = \sum \mu_i \cdot \omega_i$	μ_i: Módulo de cisalhamento do elemento *i* ω_i: Fração molar do elemento *i* na composição
Shear _std_devs_w Desvio padrão ponderado do módulo de cisalhamento	$\mu_{std\_devs\_w} = \sqrt{\sum (\mu_i - \bar{\mu})^2 \cdot \omega_i}$	μ_i: Módulo de cisalhamento do elemento *i* $\bar{\mu}$*:* Média ponderada do módulo de cisalhamento ω_i: Fração molar do elemento *i* na composição
Number_avgs_w Média ponderada do número atômico	$Z_{avgs\_w} = \sum Z_i \cdot \omega_i$	Z_i: Número atômico do elemento *i* ω_i: Fração molar do elemento *i* na composição
Number _std_devs_w Desvio padrão ponderado do número atômico	$Z_{std\_devs\_w} = \sqrt{\sum (Z_i - \bar{Z})^2 \cdot \omega_i}$	Z_i: Número atômico do elemento *i* $\bar{Z}$*:* Média ponderada do número atômico ω_i: Fração molar do elemento *i* na composição
MendeleevNumber _avgs_w Média ponderada do número de Mendeleev	$MN_{avgs\_w} = \sum MN_i \cdot \omega_i$	MN_i: Número de Mendeleev do elemento *i* ω_i: Fração molar do elemento *i* na composição
MendeleevNumber _std_devs_w Desvio padrão ponderado do número de Mendeleev	$MN_{std\_devs\_w} = \sqrt{\sum (MN_i - \overline{MN})^2 \cdot \omega_i}$	MN_i: Número de Mendeleev do elemento *i* $\overline{MN}$*:* Média ponderada do número de Mendeleev ω_i: Fração molar do elemento *i* na composição

Row_avgs_w Média ponderada do período da tabela periódica	$R_{avgs\_w} = \sum R_i \cdot \omega_i$	R_i: Período da tabela periódica do elemento i ω_i: Fração molar do elemento *i* na composição
Row _std_devs_w Desvio padrão ponderado do período da tabela periódica	$R_{std\_devs\_w} = \sqrt{\sum (R_i - \bar{R})^2 \cdot \omega_i}$	R_i: Período da tabela periódica do elemento i $\bar{R}$: Média ponderada do período da tabela periódica ω_i: Fração molar do elemento *i* na composição
AtomicRadius_avgs_w Média ponderada do raio atômico	$r_{avgs\_w} = \sum r_i \cdot \omega_i$	r_i: Raio atômico do elemento *i* ω_i: Fração molar do elemento *i* na composição
AtomicRadius _std_devs_w Desvio padrão ponderado do raio atômico	$r_{std\_devs\_w} = \sqrt{\sum (r_i - \bar{r})^2 \cdot \omega_i}$	r_i: Raio atômico do elemento *i* $\bar{r}$: Média ponderada do raio atômico ω_i: Fração molar do elemento *i* na composição
MeltT_avgs_w Média ponderada da temperatura de fusão	$T_{m\_avgs\_w} = \sum T_{m\_i} \cdot \omega_i$	T_{m_i}: Temperatura de fusão do elemento *i* ω_i: Fração molar do elemento *i* na composição
MeltT _std_devs_w Desvio padrão ponderado da temperatura de fusão	$T_{m\_std\_devs\_w} = \sqrt{\sum \left(T_{m\_i} - \overline{T_m}\right)^2 \cdot \omega_i}$	T_{m_i}: Temperatura de fusão do elemento *i* $\overline{T_m}$: Média ponderada da temperatura de fusão ω_i: Fração molar do elemento *i* na composição
BoilingT_avgs_w Média ponderada da temperatura de ebulição	$T_{b\_avgs\_w} = \sum T_{b\_i} \cdot \omega_i$	T_{b_i}: Temperatura de ebulição do elemento *i* ω_i: Fração molar do elemento *i* na composição

BoilingT _std_devs_w Desvio padrão ponderado da temperatura de ebulição	$T_{b\_std\_devs\_w} = \sqrt{\sum \left(T_{b\_i} - \overline{T_b}\right)^2 \cdot \omega_i}$	T_{b_i}: Temperatura de ebulição do elemento *i* $\overline{T_b}$: Média ponderada da temperatura de ebulição ω_i: Fração molar do elemento *i* na composição
MolarVolume_avgs_w Média ponderada do volume molar	$MV_{avgs\_w} = \sum MV_i \cdot \omega_i$	MV_i: Volume molar do elemento *i* ω_i: Fração molar do elemento *i* na composição
MolarVolume _std_devs_w Desvio padrão ponderado do volume molar	$MV_{std\_devs\_w} = \sqrt{\sum (MV_i - \overline{MV})^2 \cdot \omega_i}$	MV_i: Volume molar do elemento *i* $\overline{MV}$: Média ponderada do volume molar ω_i: Fração molar do elemento *i* na composição
AtomicSizeMismatchs Desajuste de tamanho atômico	$Mismatch = 100 \cdot \sqrt{\sum \left(1 - \frac{r_i}{\bar{r}}\right)^2 \cdot \omega_i}$	r_i: Raio atômico do elemento *i* $\bar{r}$: Média ponderada do raio atômico ω_i: Fração molar do elemento *i* na composição
Hmix* Entalpia de mistura	$H_{mix} = \sum x_a\, x_b\, (a_0 + a_1\,(x_a - x_b) + a_2\,(x_a - x_b)^2 + a_3\,(x_a - x_b)^3)$	x_a, x_b : Frações molares dos elementos *a* e *b* na composição a_0, a_1, a_2, a_3 : Coeficientes empíricos do sistema binário A-B
exCp* Capacidade térmica de excesso	$exCp = \sum x_a\, x_b\, (a_0 + a_1\,(x_a - x_b) + a_2\,(x_a - x_b)^2 + a_3\,(x_a - x_b)^3)$	x_a, x_b : Frações molares dos elementos *a* e *b* na composição a_0, a_1, a_2, a_3 : Coeficientes empíricos do sistema binário A-B
Sid Entropia de mistura ideal	$S_{id} = -\sum \omega_i \cdot \ln(\omega_i)$	ω_i: Fração molar do elemento *i* na composição

S Entropia combinada	$S = S_{id} + \sum S_{CFC\_900\_i} \cdot \omega_i$	S_{id}: Entropia de mistura ideal $S_{CFC\_900\_i}$: Entropia do elemento puro na estrutura CFC a 900°C ω_i: Fração molar do elemento *i* na composição
Cp Entropia combinada	$Cp = exCp + \sum Cp_{CFC\_900\_i} \cdot \omega_i$	$exCp$: Capacidade térmica de excesso $Cp_{CFC\_900\_i}$: Capacidade térmica específica do elemento puro na estrutura CFC a 900 °C ω_i: Fração molar do elemento *i* na composição

* Os descritores Hmix e exCp estão apresentados em sua forma simplificada, considerando uma liga composta por dois elementos. Para sistemas de ordens maiores, consultar a versão estentida em [76].

APÊNDICE B

Tabela B.1 - Lista completa dos descritores selecionados para o modelo SVM para previsão de fase, juntamente com o valor da acurácia e os valores otimizados de hiperparâmetros (*C*, *kernel*, *degree*, *gamma*).

Descitores	**Acurácia**	***C***	***kernel***	***degree***	***gamma***
VEC_avgs_w	0,73	0,67	rbf	False	scale
MendeleevNumber_avgs_w	0,80	0,64	rbf	False	scale
Electronegativity_std_devs_w	0,82	21,00	rbf	False	auto
Hfus_std_devs_w	0,85	978,82	rbf	False	auto
Shear_avgs_w	0,91	985,98	rbf	False	auto
Hfus_avgs_w	0,95	965,81	rbf	False	scale
Sid	0,96	998,92	rbf	False	auto
Row_std_devs_w	0,97	997,56	rbf	False	auto
Cp	0,97	954,53	rbf	False	scale
Sfus_std_devs_w	0,97	916,65	rbf	False	scale
Shear_std_devs_w	0,97	920,08	rbf	False	auto
VEC_std_devs_w	0,97	978,08	rbf	False	auto
BoilingT_avgs_w	0,97	955,44	rbf	False	scale
Column_avgs_w	0,97	245,64	rbf	False	auto
Sfus_avgs_w	0,97	921,55	rbf	False	auto
MolarVolume_std_devs_w	0,97	907,49	rbf	False	scale
BoilingT_std_devs_w	0,97	972,28	rbf	False	scale
AtomicWeight_std_devs_w	0,97	974,71	rbf	False	scale
Density_avgs_w	0,97	386,08	rbf	False	scale
MeltT_std_devs_w	0,97	931,48	rbf	False	auto
Number_std_devs_w	0,97	996,73	rbf	False	scale
AtomicSizeMismatchs	0,97	988,93	rbf	False	scale
S	0,97	987,02	rbf	False	auto
MolarVolume_avgs_w	0,97	991,50	rbf	False	auto
Column_std_devs_w	0,97	971,55	rbf	False	scale
Density_std_devs_w	0,97	986,20	rbf	False	scale
exCp_l	0,97	933,55	rbf	False	auto
AtomicRadius_std_devs_w	0,97	960,92	rbf	False	auto
MendeleevNumber_std_devs_w	0,97	939,94	rbf	False	auto
Number_avgs_w	0,97	989,91	poly	4	auto
Electronegativity_avgs_w	0,97	994,25	rbf	False	scale
AtomicRadius_avgs_w	0,97	996,27	rbf	False	scale
Hmix_l	0,97	965,27	poly	3	auto
AtomicWeight_avgs_w	0,96	999,80	rbf	False	scale
MeltT_avgs_w	0,97	963,41	rbf	False	scale
Row_avgs_w	0,97	997,93	rbf	False	auto

APÊNDICE C

Tabela C.1 - Lista completa dos descritores selecionados para o modelo SVR para previsão da energia de falha de empilhamento, juntamente com parâmetros para medir a robustez do modelo (MAE, MSE, RMSE, R2) e os valores otimizados dos hiperparâmetros (*C, kernel, épsilon, gamma*).

Descritores	**MAE**	**MSE**	**RMSE**	**R2**	***C***	***kernel***	***epsilon***	***gamma***
Hmix_l	54,79	5599,93	74,83	0,24	158,76	rbf	0,0111	auto
Density_std_devs_w	39,99	3197,08	56,54	0,57	626,20	rbf	0,0200	auto
Sfus_avgs_w	33,16	2781,73	52,74	0,62	997,89	rbf	0,4063	scale
BoilingT_std_devs_w	19,57	1253,81	35,41	0,83	464,93	rbf	0,1168	auto
MendeleevNumber_avgs_w	18,46	1064,82	32,63	0,86	644,58	rbf	0,0014	scale
Density_avgs_w	15,69	867,27	29,45	0,88	819,25	rbf	0,1812	auto
Electronegativity_avgs_w	16,10	871,87	29,53	0,88	994,94	rbf	0,0018	auto
Electronegativity_std_devs_w	15,84	891,15	29,85	0,88	980,94	rbf	0,0038	auto
Sfus_std_devs_w	14,97	887,88	29,80	0,88	989,98	rbf	0,7736	auto
AtomicRadius_avgs_w	14,59	867,53	29,45	0,88	926,21	rbf	0,7052	auto
Column_avgs_w	14,81	847,90	29,12	0,89	994,89	rbf	0,1994	auto
AtomicWeight_avgs_w	15,34	853,59	29,22	0,88	999,70	rbf	0,0247	auto
Sid	14,78	850,43	29,16	0,89	729,48	rbf	0,0521	auto
VEC_avgs_w	14,44	816,99	28,58	0,89	665,79	rbf	0,0255	auto
Number_std_devs_w	14,53	819,38	28,62	0,89	701,05	rbf	0,0139	auto
MolarVolume_std_devs_w	14,43	797,84	28,25	0,89	720,13	rbf	0,0020	scale
exCp_l	14,84	811,57	28,49	0,89	780,25	rbf	0,0050	scale
MolarVolume_avgs_w	14,91	797,65	28,24	0,89	756,60	rbf	0,0506	auto
Number_avgs_w	14,91	795,60	28,21	0,89	716,78	rbf	0,0055	scale
Cp	15,14	863,83	29,39	0,88	636,76	rbf	0,0257	auto
BoilingT_avgs_w	15,03	850,29	29,16	0,89	567,73	rbf	0,0010	auto
VEC_std_devs_w	14,82	837,76	28,94	0,89	499,28	rbf	0,0092	auto
S	14,64	822,65	28,68	0,89	469,95	rbf	0,0036	scale
Shear_avgs_w	14,99	848,10	29,12	0,89	446,13	rbf	0,0059	auto
MeltT_avgs_w	15,16	853,18	29,21	0,88	417,09	rbf	0,0239	auto
MendeleevNumber_std_devs_w	15,48	861,97	29,36	0,88	390,18	rbf	0,0018	auto
Shear_std_devs_w	15,62	865,02	29,41	0,88	437,74	rbf	0,0057	scale
Column_std_devs_w	15,93	883,89	29,73	0,88	453,85	rbf	0,0491	scale
Hfus_std_devs_w	16,30	892,99	29,88	0,88	390,54	rbf	0,0019	scale
Hfus_avgs_w	16,28	897,37	29,96	0,88	398,15	rbf	0,0850	auto
MeltT_std_devs_w	16,32	904,30	30,07	0,88	418,37	rbf	0,0017	scale
AtomicWeight_std_devs_w	16,48	919,27	30,32	0,88	422,62	rbf	0,0495	scale
AtomicRadius_std_devs_w	16,57	928,98	30,48	0,87	445,67	rbf	0,0012	scale
AtomicSizeMismatchs	16,73	938,03	30,63	0,87	476,95	rbf	0,0018	scale
Row_avgs_w	16,86	949,23	30,81	0,87	491,91	rbf	0,2351	scale
Row_std_devs_w	16,98	965,20	31,07	0,87	488,37	rbf	0,5765	scale