MACHINE LEARNING REPRESENTATION FOR ATOMIC ENERGIES IN MAGNETIC MATERIALS

Nguyen Tien Cuong[1*], Nguyen Viet Anh[1], Nguyen Truong Danh[1], Pham Tien Lam[2]
[1]VNU - University of Science, Hanoi, [2]Phenikaa University

ARTICLE INFO

Received: 27/6/2024
Revised: 30/9/2024
Published: 30/9/2024

KEYWORDS

Machine learning
Linear regression
LASSO regression
Rigde regression
Atomic energies
Magnetic materials

ABSTRACT

In this study, we propose machine learning models, including linear regression, LASSO regression, and Ridge regression, for fast estimating atomic energies in a magnetic system. In our method, the total energy of a magnetic system contains chemical energy and magnetic energy. The chemical energy is approximated as the summation of atomic energy which is the interaction energy with its surrounding chemical environment within a certain cutoff radius. Atomic energy is decomposed into two-body terms which are expressed as a linear combination of basis functions. The magnetic energy is also approximated as the summation of atomic magnetic energy. The machine learning models, trained with crystal bcc-Fe data, can fast estimate the total energy of the system in both magnetic and non-magnetic states. Result from these models were analyzed and compared with calculated results by density functional theory (DFT). Model evaluation metrics including MSE, MAE and R^2 indicated that Ridge regression gives the best results. Results from our machine learning models show good agreement with DFT calculations.

CÁC MÔ HÌNH HỌC MÁY BIỂU DIỄN NĂNG LƯỢNG NGUYÊN TỬ TRONG CÁC VẬT LIỆU TỪ

Nguyễn Tiến Cường[1*], Nguyễn Việt Anh[1], Nguyễn Trường Danh[1], Phạm Tiến Lâm[2]
[1]Trường Đại học Khoa học Tự nhiên - ĐH Quốc gia Hà Nội, [2]Trường Đại học Phenikaa

THÔNG TIN BÀI BÁO

Ngày nhận bài: 27/6/2024
Ngày hoàn thiện: 30/9/2024
Ngày đăng: 30/9/2024

TỪ KHÓA

Học máy
Hồi quy tuyến tính
Hồi quy LASSO
Hồi quy Rigde
Năng lượng nguyên tử
Vật liệu từ

TÓM TẮT

Trong nghiên cứu này, chúng tôi đề xuất các mô hình học máy, bao gồm hồi quy tuyến tính, hồi quy LASSO và hồi quy Rigde, để ước tính nhanh năng lượng tổng cộng của các hệ vật liệu từ. Trong phương pháp của chúng tôi, năng lượng của một hệ vật liệu từ là tổng của năng lượng tương tác hóa học và năng lượng tương tác từ. Năng lượng tương tác hóa học của hệ được tính gần đúng như là tổng của các năng lượng nguyên tử cấu thành, khi tương tác với môi trường hóa học xung quanh trong một bán kính giới hạn xác định. Năng lượng của từng nguyên tử được phân tách thành các số hạng tương tác hai vật và biểu diễn dưới dạng tổ hợp tuyến tính của các hàm cơ sở. Năng lượng tương tác từ cũng được tính gần đúng như là tổng năng lượng tương tác từ của các nguyên tử cấu thành. Các mô hình học máy, sau khi được huấn luyện với dữ liệu của mạng tinh thể bcc-Fe, có thể dự đoán nhanh năng lượng tổng cộng của hệ ở cả trạng thái có và không có từ tính. Kết quả từ các mô hình này đã được phân tích và so sánh với kết quả tính toán bằng lý thuyết phiếm hàm mật độ (DFT). Các chỉ số đánh giá mô hình như MSE, MAE và R^2 chỉ ra rằng mô hình hồi quy Ridge cho kết quả tốt nhất. Kết quả tính toán từ các mô hình học máy của chúng tôi cho thấy sự phù hợp tốt với các tính toán DFT.

DOI: https://doi.org/10.34238/tnu-jst.10668

[*] Corresponding author. *Email: ntcuong@hus.edu.vn*

1. Giới thiệu

Trong những năm gần đây việc ứng dụng trí truệ nhân tạo, cụ thể hơn là học máy vào lĩnh vực khoa học vật liệu đã và đang thu hút được sự chú ý của nhiều nhà khoa học trong và ngoài nước. Việc ứng dụng học máy được kì vọng sẽ giúp tăng tốc quá trình nghiên cứu, thiết kế vật liệu mới.

Tính toán năng lượng của một hệ vật liệu đóng vai trò quan trọng trong việc xác định cấu trúc điện tử và các tính chất của hệ vật liệu đó. Việc mô hình hóa đúng về năng lượng sẽ giúp ích chúng ta trong việc tính toán, mô phỏng vật liệu. Năng lượng của một hệ vật liệu có thể nhận được bằng cách thực hiện tính toán cấu trúc điện tử dựa trên lý thuyết phiếm hàm mật độ (Density Functional Theory-DFT) [1], [2]. Hiện nay, các tính toán DFT được coi là một trong các phương pháp tính toán chuẩn mực, đáng tin cậy và thường được dùng làm tham chiếu cho các phương pháp tính toán khác. Tuy nhiên tính toán DFT cho các hệ vật liệu lớn gồm nhiều nguyên tử đòi hỏi hiệu năng máy tính cao và thời gian tính toán kéo dài. Vì vậy cần phải có phương pháp giúp giảm thiểu thời gian và đòi hỏi ít chi phí tính toán hơn.

Thông thường, bề mặt thế năng (Potential Energy Surface-PES) của hệ được xây dựng dưới dạng tổng từ các đóng góp của các số hạng thấp chiều đơn giản (các yếu tố cấu trúc) biểu thị các liên kết: cộng hóa trị (covalent bonds), liên kết góc (bond angles) và góc nhị diện (dihedral angles) [3]. Các phương pháp này tỏ ra hiệu quả và được áp dụng rộng rãi để mô phỏng các hệ sinh học lớn (large biosystem). Nhưng chúng khó có thể mô tả các phản ứng hóa học (chemical reactions) liên quan đến sự hình thành hoặc sự phân ly của các liên kết cộng hóa trị. Gần đây, các phương pháp thay thế, có thể "học" PES từ bộ dữ liệu lớn về các cấu trúc vật liệu và năng lượng DFT tương ứng, đã và đang được phát triển mạnh mẽ [4] – [10].

Trong nghiên cứu trước đây, nhóm chúng tôi đã thành công trong việc phát triển các mô hình mạng Nơ-ron nhân tạo có thể "học" được các tính chất hóa học, tính chất vật lý ẩn trong các hệ vật liệu dựa trên các dữ liệu vật liệu đã biết [11]. Ngoài ra, chúng tôi đã thành công trong việc phát triển các mô hình hồi quy tuyến tính, các mô hình học sâu trong biểu diễn tương tác cặp (pairwise interactions) cho bề mặt thế năng, lực nguyên tử và năng lượng trong các hệ vật liệu không từ tính, như hệ Silic tinh thể và vô định hình [12], [13]. Với các hệ vật liệu từ tính, việc đề xuất mô hình ước tính nhanh năng lượng phức tạp hơn nhiều so với các hệ không từ tính. Ngoài ra, việc chuẩn bị và chuẩn hóa dữ liệu tính toán DFT cho các vật liệu từ, dùng để huấn luyện mô hình cũng gặp nhiều khó khăn.

Trong bài báo này, chúng tôi đề xuất các mô hình học máy dựa trên hồi quy tuyến tính nhằm ước tính nhanh năng lượng tổng cộng của các hệ vật liệu từ với độ chính xác cao. Trong hồi quy tuyến tính các hệ số không bị ràng buộc nên có thể khớp tốt cho tập dữ liệu huấn luyện, nhưng khả năng dự báo cho những điểm dữ liệu mới không được tốt. Hiện tượng này trong học máy được gọi là hiện tượng "quá khớp" (overfitting), đây là vấn đề rất quan trọng đối với học máy. Nhằm kiểm soát hiện tượng quá khớp, chúng tôi thử nghiệm thêm hồi quy Ridge và hồi quy LASSO (Least Absolute Shrinkage and Selection Operator). Trong hồi quy Ridge, các hệ số hồi quy tuyến tính được thừa nhận là các biến ngẫu nhiên độc lập tuân theo phân phối chuẩn với kỳ vọng toán là 0 và độ lệch chuẩn là siêu tham số (hyperparameter) để kiểm soát khả năng khái quát hoá của mô hình. Tương tự như vậy trong hồi quy LASSO, chúng ta thừa nhận các hệ số tuân theo phân phối Laplace. Chúng tôi đã tiến hành mô hình hóa, lấy mẫu và tự chạy các tính toán DFT để chuẩn bị dữ liệu huấn luyện cho các mô hình này.

2. Mô hình hóa và chuẩn bị dữ liệu

2.1. Các mô hình hồi quy

Hồi quy tuyến tính (Linear Regression) hay phương pháp bình phương tối thiểu (least square) là phương pháp hồi quy đơn giản và cổ điển nhất. Đối với mô hình này, giá trị dự đoán tại điểm dữ liệu thứ *i* sẽ có dạng đơn giản như sau:

$$\hat{Y}_i = \sum_{j=1}^{p} w_j x_{ij} \quad (1)$$

Để xác định *wj*, chúng ta cực tiểu hóa hàm mất mát *L(wj)*:

$$L(w_j) = \frac{1}{2}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 = \frac{1}{2}\sum_{i=1}^{N}(Y_i - \sum_{j=1}^{p} w_j x_{ij})^2 \quad (2)$$

Hồi quy tuyến tính có ưu điểm là mô hình đơn giản dễ triển khai. Tuy nhiên, khi làm việc với các hệ phức tạp với lượng dữ liệu lớn, thì kết quả dự đoán của mô hình thường cho sai số khá lớn.

Để cải thiện nhược điểm của mô hình hồi quy tuyến tính, người ta đưa ra mô hình hồi quy LASSO. Hồi quy LASSO tối ưu hóa tham số theo quy tắc L1 (L1 regularization). Mô hình thêm vào tham số β (tham số penalty), là một số dương, giúp chúng ta có thể điều chỉnh để tăng cường khả năng khái quát hoá của mô hình cho việc dự đoán các điểm dữ liệu mới:

$$L(w_j) = \frac{1}{2}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 = \frac{1}{2}\sum_{i=1}^{N}(Y_i - \sum_{j=1}^{p} w_j x_{ij})^2 + \beta \sum_{j=1}^{p}|w_j| \quad (3)$$

Tương tự như hồi quy LASSO, hồi quy Ridge tối ưu hóa tham số theo quy tắc L2 (L2 regularization). Tuy nhiên, hàm mất mát có sự khác biệt ở số hạng cuối:

$$L(w_j) = \frac{1}{2}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 = \frac{1}{2}\sum_{i=1}^{N}(Y_i - \sum_{j=1}^{p} w_j x_{ij})^2 + \beta \sum_{j=1}^{p} {w_j}^2 \quad (4)$$

Giá trị β càng lớn thì mức độ ràng buộc để cho các hệ số hồi quy tiến tới không càng mạnh, ngược lại nếu β bằng không thì mô hình hồi quy Ridge và mô hình hồi quy LASSO suy biến thành hồi quy tuyến tính. Trong công trình này, chúng tôi sử dụng phương pháp Grid search kết hợp với 5-fold cross validation để tìm kiếm giá trị β tối ưu. Cụ thể, hàm GridSearchCV của thư viện Scikit-learn được sử dụng để tìm kiếm β [14].

Các mô hình hồi quy nêu trên đều có những ưu nhược điểm riêng. Việc lựa chọn mô hình nào tùy thuộc vào từng bài toán cụ thể. Trong nghiên cứu này, chúng tôi khảo sát cả ba mô hình để tìm ra mô hình có thể biểu diễn tốt nhất năng lượng nguyên tử trong các hệ vật liệu từ.

2.2. Biểu diễn năng lượng tương tác giữa các nguyên tử

Năng lượng của các nguyên tử trong các hệ vật liệu từ tính chịu ảnh hưởng của hai thành phần: năng lượng tương tác hóa học, E^c và năng lượng tương tác từ, E^{mag}. Năng lượng tổng cộng của hệ sẽ là tổng đóng góp từ năng lượng của cả hai thành phần.

$$E = E^c + E^{mag} \quad (5)$$

Trong mô hình, tổng năng lượng của một hệ sẽ được tính gần đúng bằng tổng của tất cả các năng lượng của các nguyên tử cấu thành được xác định bởi tương tác giữa nguyên tử được chọn với các nguyên tử lân cận trong môi trường hóa học xung quanh trong một bán kính giới hạn (cutoff energy). Năng lượng nguyên tử sẽ được biểu thị dưới dạng tổ hợp tuyến tính của các hàm cơ sở, sử dụng các phương pháp hồi quy (Linear, LASSO và Rigde) để tối ưu tham số của mô hình này.

Giả định rằng tổng năng lượng của một hệ là tổng năng lượng tương tác của các nguyên tử cấu thành: $E^c = \sum_i E_i^c$, trong đó E^c là tổng năng lượng và E_i^c là năng lượng đóng góp của nguyên tử thứ i vào tổng năng lượng. Các năng lượng của nguyên tử i có thể được biểu diễn dựa trên tương tác của nó với các nguyên tử lân cận trong một bán kính giới hạn, r_c. Trong mô hình của chúng tôi chỉ tính đến các số hạng tương tác hai vật (two-body interactions). Về nguyên tắc, các số hạng tương tác bậc cao hơn có thể được đưa vào để nâng cao hiệu quả của các mô hình học máy.

$$E_i^c = \sum_j E_{ij} = \sum_j \sum_k c_k b_k(r_{ij}, \vartheta_i, \vartheta_j) = \sum_k c_k \sum_j b_k(r_{ij}, \vartheta_i, \vartheta_j) = \sum_k c_k x_k \quad (6)$$

Trong đó $x_k = \sum_j b_k(r_{ij}, \vartheta_i, \vartheta_j)$, $b_k(r_{ij}, \vartheta_i, \vartheta_j)$ là các hàm cơ sở, c_k là hệ số khai triển, ϑ_i và ϑ_j là các vectơ đặc trưng mã hóa thông tin của nguyên tử i và j, tương ứng. Để nâng cao hiệu quả biểu diễn phi tuyến, các số hạng tương tác hai vật được khải triển theo đa thức đến bậc *p*:

$$E_i^c = \sum_k c_k^{(1)} x_k + \sum_k c_k^{(2)} x_k^2 + \cdots + \sum_k c_k^{(p)} x_k^p \quad (7)$$

Khi áp dụng cho hệ vật liệu từ (ví dụ hệ bcc-Fe), vì thông tin nguyên tử giống hệt nhau đối với tất cả các nguyên tử sắt, chúng ta có thể loại bỏ ϑ_i và ϑ_j trong phương trình (6). Do đó (6) có thể được rút gọn thành dạng đơn giản hơn:

$$E_i^c = \sum_j \sum_k c_k b(r_{ij})_k \quad (8)$$

Chúng tôi sử dụng dạng biểu diễn hàm cơ sở $b(r_{ij})$=$e^{-\eta_k(r-r_k)^2} f_c(r_{ij})$ tương ứng cho hàm cơ sở Gaussian. Trong đó η_k, r_k là các tham số xác định các hàm cơ sở, $f_c(r_{ij})$ là một hàm cắt (hàm giới hạn) đảm bảo rằng năng lượng thay đổi một cách liên tục và trơn ở bán kính cắt, r_c:

$$f_c(r_{ij}) = \begin{cases} 0{,}5[\cos\left(\frac{\pi r_{ij}}{r_c}\right) + 1] \; nếu \;\; r_{ij} < r_c \\ 0 \qquad\qquad\qquad nếu \;\; r_{ij} > r_c \end{cases} \quad (9)$$

Hệ số c_k được xác định bằng một phép hồi quy tuyến tính theo năng lượng được tính từ DFT.

Đối với số hạng tương tác từ, chúng ta giả định rằng tổng năng lượng tương tác từ của hệ là tổng năng lượng tương tác từ của các nguyên tử cấu thành: $E^{mag} = \sum_i E_i^{mag}$, trong đó E^{mag} là tổng năng lượng và E_i^{mag} là năng lượng đóng góp của nguyên tử thứ i vào tổng năng lượng tương tác từ.

Năng lượng tương tác từ được tính theo công thức:

$$E_i^{mag} = f_1(\sigma_i) + \sum_j f_2(\sigma_i, \sigma_j, r_{ij}) f_c(r_{ij}) + \sum_{jk} f_3(\sigma_i, \sigma_j, \sigma_k, r_{ij}, r_{ik}, r_{jk}) f_c(r_{ij}) f_c(r_{ik}) f_c(r_{jk}) \quad (10)$$

Trong các mô hình Ising, năng lượng được biểu diễn đơn giản hóa như sau:

$$E_i^{mag} = A\sigma_i + \sum_j \sigma_i \sigma_j f_2(r_{ij}) f_c(r_{ij}) + \sum_{jk} \sigma_i \sigma_j \sigma_k f_3(r_{ij}, r_{ik}, r_{jk}) f_c(r_{ij}) f_c(r_{ik}) f_c(r_{jk}) \quad (11)$$

Trong đó: $$f(r_{ij}) = \sum_\gamma c_\gamma b_\gamma(r_{ij}) \quad (12)$$

Và $$f(r_{ij}, r_{ik}, r_{jk}) = \sum_\alpha c_\alpha b_\alpha(r_{ij}, r_{ik}, r_{jk}) \quad (13)$$

Cuối cùng, năng lượng tổng cộng của một hệ vật liệu từ được tính theo công thức:

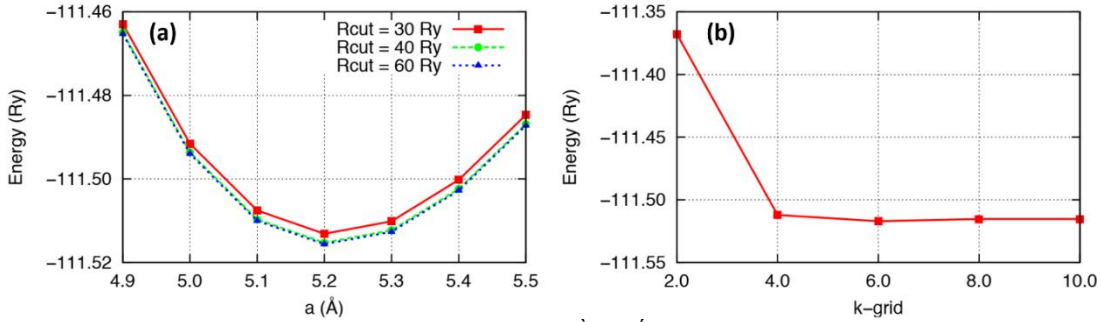
$$E_{total} = \sum_i E_i = \sum_i (E_i^c + E_i^{mag}) = \sum_i E_i^c + \sum_i E_i^{mag} \quad (14)$$

2.3. Chuẩn bị dữ liệu

Với mô hình ước tính năng lượng tương tác hóa học, chúng tôi sử dụng mạng tinh thể 3x3x3 cho tinh thể bcc-Fe với 54 nguyên tử và 5872 cấu trúc được tạo ngẫu nhiên. Năng lượng và lực của cấu trúc tinh thể được tính toán bằng PWscf code trong quantum ESPRESSO. Hàm PBE được sử dụng để biểu diễn năng lượng tương quan trao đổi. Các tham số mô hình được lựa chọn dựa trên kết quả cực tiểu hóa năng lượng, giá trị tham số mô hình được lựa chọn khi xu hướng giá trị năng lượng dần ổn định. Hình 1a và 1b lần lượt mô tả sự phụ thuộc của năng lượng vào hằng số mạng tại các giá trị năng lượng giới hạn khác nhau và vào lưới chia *k-grid*.

Để tối ưu hóa năng lượng cutoff, bán kính cutoff (Rcut) được thay đổi lần lượt với các giá trị là 30Ry, 40Ry và 60Ry. Từ hình 1a, chúng ta thấy rằng với giá trị năng lượng Rcut từ 40Ry thì năng lượng có xu hướng ổn định. Vì vậy, tham số năng lượng cutoff với Rcut = 40Ry được lựa chọn.

Để tối ưu hóa lưới chia *k-grid*, lưới chia được thay đổi lần lượt tại các giá trị: kxkxk lần lượt là 2x2x2, 4x4x4, 6x6x6, 8x8x8 và 10x10x10. Từ hình 1b, chúng ta thấy rằng với giá trị k ≥ 4 thì năng lượng có xu hướng ổn định. Vì vậy, tham số *k-grid* = 4x4x4 được lựa chọn cho toàn bộ các tính toán DFT để chuẩn bị dữ liệu cho mô hình.

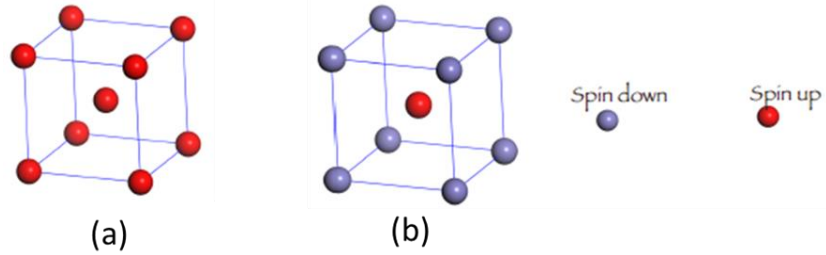


Hình 1. *Sự phụ thuộc của năng lượng vào: (a) hằng số mạng tại các giá trị năng lượng giới hạn khác nhau và (b) vào lưới chia k-grid*

Với mô hình ước tính năng lượng tương tác từ, chúng tôi giả thiết rằng tọa độ nguyên tử trong mạng tinh thể lý tưởng được giữ cố định, cấu hình spin được tạo ra một cách ngẫu nhiên và tạo ra bộ dữ liệu gồm 1497 cấu trúc trong không gian spin được tham khảo từ các nghiên cứu trước đây về bcc-Fe [15]-[17]. Trong không gian spin: D = {(spin, E_i), i=1...m}, chúng ta cần điều khiển spin trong quá trình tính tự hợp SCF để cực tiểu hóa năng lượng với điều kiện mô men spin ở các sites cố định:

$$E = E_{SCF} + \lambda \sum_i (\hat{\mu}_i - \mu_i)^2 \quad (15)$$

Để lựa chọn giá trị tham số λ trong công thức (15), chúng tôi xét cả hai mô hình cấu trúc tinh thể: Mô hình cấu trúc tinh thể cho trạng thái sắt từ (toàn bộ các nút mạng là spin- up, như hình 2a) và mô hình cấu trúc tinh thể cho trạng thái phản sắt từ (các nút mạng ở đỉnh là spin – down và nút trung tâm là spin – up, như hình 2b).



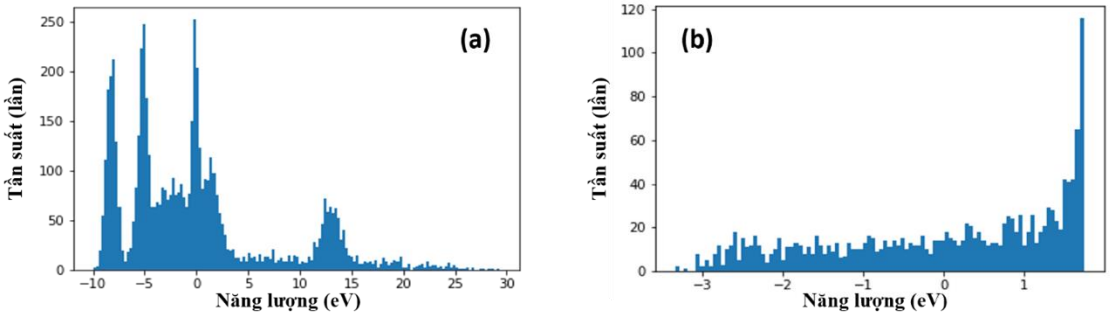
Hình 2. *Cấu hình spin cho: (a) trạng thái sắt từ và (b) trạng thái phản sắt từ*

Các kết quả tính toán sự phụ thuộc của năng lượng và độ từ hóa vào giá trị giới hạn mô men từ với các giá trị tham số λ khá nhau của chúng tôi chỉ ra rằng khi giá trị tham số $\lambda \geq 20$ thì hệ từ tính đạt trạng thái ổn định, trong cả hai trường hợp sắt từ và phản sắt từ. Vì vậy, tham số λ = 20 được lựa chọn cho toàn bộ các tính toán DFT để chuẩn bị dữ liệu cho mô hình.

3. Huấn luyện mô hình, kết quả và thảo luận

3.1. Huấn luyện mô hình

Bộ dữ liệu của mạng tinh thể bcc-Fe 3x3x3 ở trạng thái không từ tính gồm 5872 cấu trúc được chia theo tỷ lệ 80-20 thành dữ liệu huấn luyện (training data) với 4697 cấu trúc và dữ liệu kiểm tra (test data) với 1175 cấu trúc, cho các mô hình ước tính năng lượng tương tác hóa học. Tương tự, bộ dữ liệu gồm 1497 cấu trúc của mạng tinh thể bcc-Fe ở trạng thái từ tính với các cấu hình spin khác nhau, được chia theo tỷ lệ 80-20 thành dữ liệu huấn luyện gồm 1197 cấu trúc và dữ liệu kiểm tra gồm 300 cấu trúc. Hình 3a và 3b lần lượt biểu diễn sự phân bố năng lượng tính bằng phương pháp DFT của các bộ dữ liệu mạng tinh thể bcc-Fe ở trạng thái không có và có từ tính.



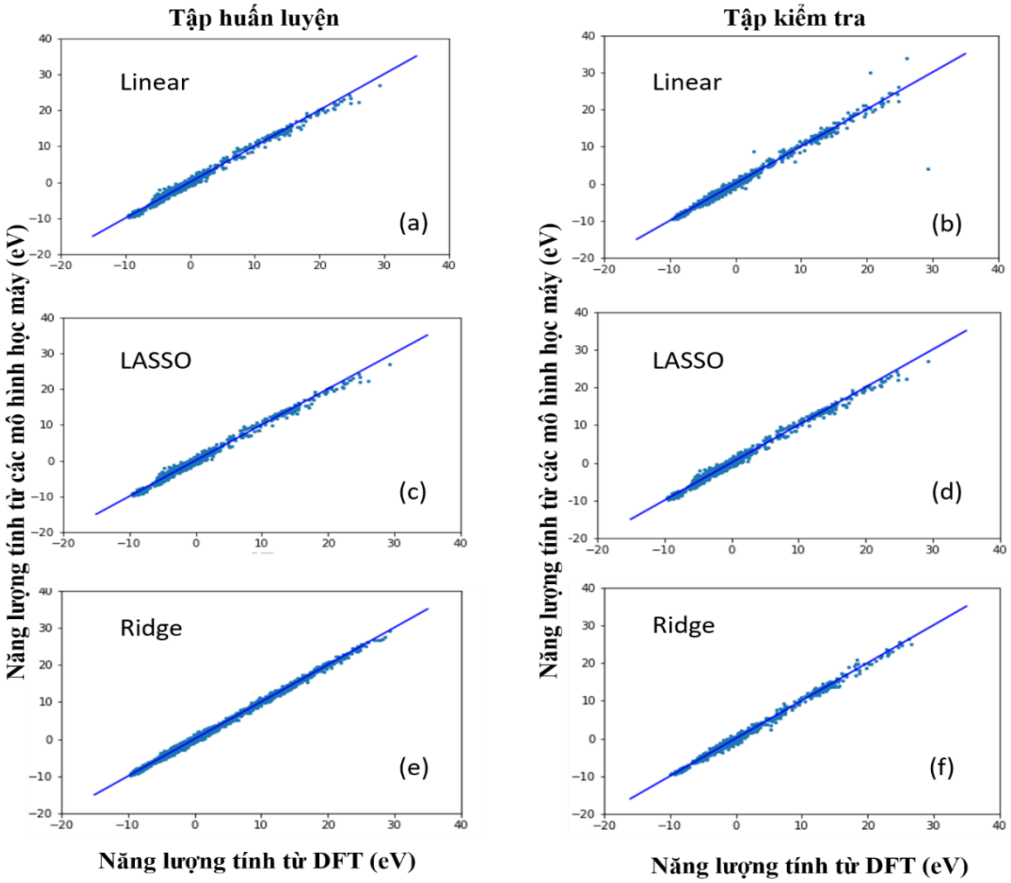
Hình 3. *Biểu đồ biểu diễn sự phân bố năng lượng tính bằng phương pháp DFT của: (a) Bộ dữ liệu ở trạng thái không từ tính, (b) Bộ dữ liệu ở trạng thái có từ tính*

Trong mô hình dự đoán năng lượng của hệ ở trạng thái không có từ tính (chỉ bao gồm tương tác hóa học), với cả ba mô hình hồi quy: Linear, LASSO và Rigde, chúng tôi sử dụng bán kính

giới hạn r_c= 6Å. Đối với các hàm Gaussian, trước tiên chúng tôi sử dụng hàm với trung tâm là nguyên tử trung tâm, r_k = 0.0. Số lượng các hàm cơ sở Gaussian được xác định bởi tham số, η_{max}. Các hàm cơ sở Gaussian với các tâm khác nhau: $r_k \in \{0.0, \ldots, 8.0\}$ và $\eta_k \in \{0.05, \ldots, 2.05\}$ cũng đã được mở rộng sau đó. Trong mô hình dự đoán năng lượng của hệ ở trạng thái có từ tính (bao gồm cả tương tác hóa học và tương tác từ), chúng tôi giữ nguyên các thông số về bán kính cắt, η_k, r_k với cả 3 phương pháp hồi quy Linear, LASSO, Ridge như trên, đồng thời bổ sung thêm các yếu tố từ tính, các mô men từ cho mô hình tính toán.

3.2. Kết quả và thảo luận

3.2.1. Dự đoán năng lượng cho hệ ở trạng thái không từ tính



Hình 4. *So sánh kết quả tính toán năng lượng của hệ bcc-Fe trong trạng thái không từ tính giữa các mô hình học máy: (a), (b) hồi quy tuyến tính; (c), (d) hồi quy LASSO và (e), (f) hồi quy Ridge, với cả tập huấn luyện và kiểm tra*

Hình 4 đưa ra so sánh kết quả tính toán năng lượng của hệ bcc-Fe trong trạng thái không từ tính giữa các mô hình học máy và DFT của cả tập huấn luyện và tập kiểm tra. Chúng ta thấy rằng đa số các điểm biểu diễn kết quả dự đoán từ các mô hình học máy của chúng tôi đều phân bố chủ yếu dọc theo đường phân giác thứ nhất với một số ít điểm bị lệch xa. Điều này chứng tỏ rằng mô hình mà chúng tôi đề xuất và huấn luyện có thể dự đoán năng lượng tương tác hóa học của các hệ vật liệu không từ tính với độ chính xác cao khi so với tính toán từ DFT.

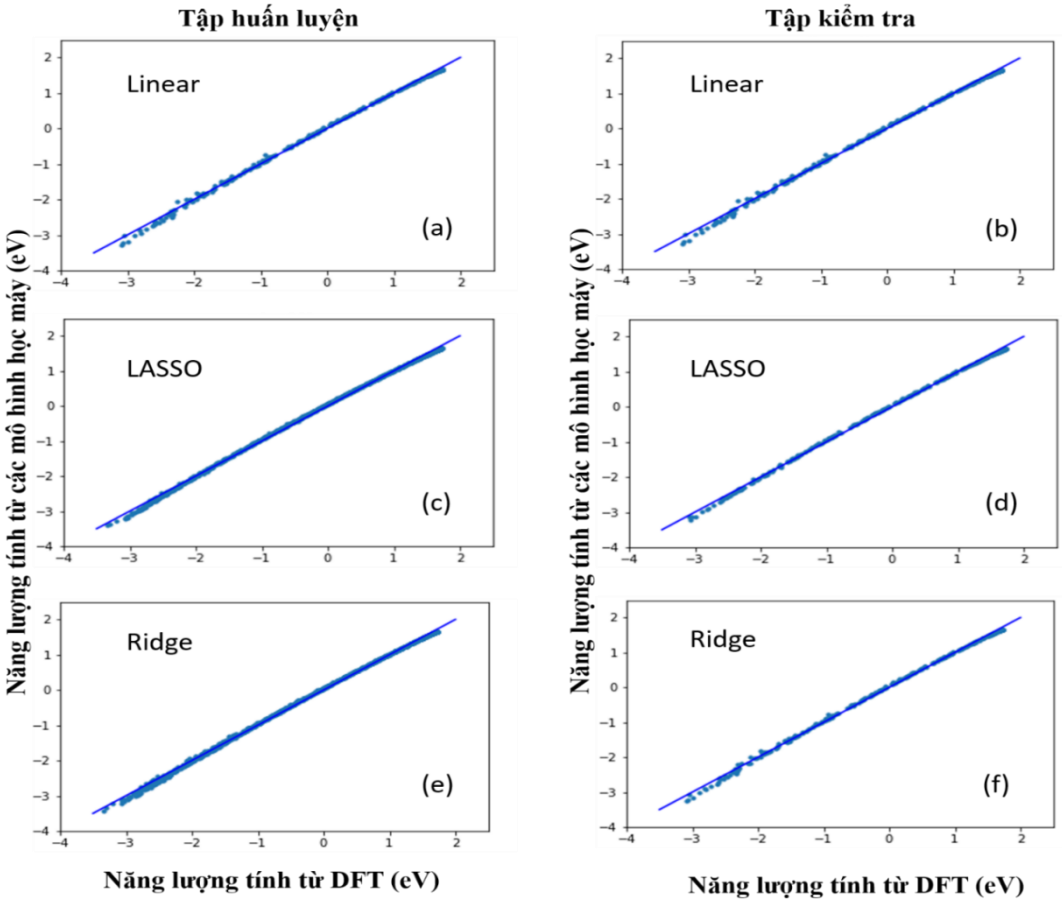
Nhằm lượng hóa, các chỉ số đánh giá sai số MSE, MAE và R^2 của các mô hình hồi quy nêu trên được liệt kê ra ở Bảng 1. Kết quả chỉ ra rằng: Mô hình sử dụng hồi quy Linear cho kết quả tính năng lượng của tập huấn luyện rất tốt (MSE ~0,077, MAE ~0,218, R^2 ~ 0,998), tuy nhiên với tập kiểm tra lại cho kết quả có độ chính xác không cao (MSE ~5,909, MAE ~0,496, R^2 ~ 0,894).

Đây là dấu hiệu của hiện tượng overfiting. Kết quả thực nghiệm của chúng tôi cũng cho thấy các mô hình sử dụng hồi quy Ridge và LASSO cho kết quả tính năng lượng kém hơn với tập huấn luyện, nhưng lại cho kết quả tốt hơn tập kiểm tra (xem bảng 1). Đây là một minh chứng rõ ràng cho thấy chúng ta có thể kiểm soát hiện tượng overfiting của mô hình hồi quy tuyến tính khi sử dụng các hệ số penalty để đưa thêm các ràng buộc ban đầu (prior) cho các hệ số hồi quy. Các chỉ số đánh giá mô hình như MSE, MAE và R^2 cho thấy mô hình hồi quy Ridge cho kết quả tốt nhất trên cả hai bộ dữ liệu huấn luyện và kiểm tra.

Bảng 1. *Đánh giá sai số so với kết quả tính toán từ DFT của các mô hình dự đoán năng lượng của hệ ở trạng thái không từ tính*

Mô hình	Tập dữ liệu	MAE (eV/nguyên tử)	RMSE (eV/ nguyên tử)	R^2
Hồi quy tuyến tính	Huấn luyện	0,077325	0,217982	0,998629
	Kiểm tra	5,909171	0,495692	0,894121
Hồi quy LASSO	Huấn luyện	0,374105	0,462868	0,993369
	Kiểm tra	0,391864	0,470368	0,992978
Hồi quy Rigde	Huấn luyện	0,117007	0,259523	0,997951
	Kiểm tra	0,225884	0,340069	0,995745

3.2.2. Dự đoán năng lượng cho hệ ở trạng thái có từ tính



Hình 5. *So sánh kết quả tính toán năng lượng của hệ bcc-Fe trong trạng thái có từ tính giữa các mô hình học máy: (a), (b) hồi quy tuyến tính; (c), (d) hồi quy LASSO và (e), (f) hồi quy Ridge, với cả tập huấn luyện và kiểm tra*

Hình 5 đưa ra so sánh kết quả tính toán năng lượng của hệ bcc-Fe trong trạng thái có từ tính giữa các mô hình học máy và DFT của cả tập huấn luyện và tập kiểm tra. Ở trạng thái có từ tính, chúng ta phải kể đến đồng thời cả hai số hạng năng lượng là tương tác hóa học và năng lượng từ. Từ kết quả đồ thị thu được, chúng ta thấy rằng các mô hình học máy đã xây dựng có thể đưa ra kết quả ước tính năng lượng tương tác có độ chính xác cao khi so với kết quả tính toán bằng DFT. Trong đó, kết quả từ mô hình hồi quy tuyến tính cho kết quả có nhiều điểm lệch khỏi đường phân giác thứ nhất hơn so với hai mô hình còn lại, chứng tỏ một cách định tính rằng phương pháp này có độ chính xác thấp hơn.

Để định lượng sai số, một cách tương tự, bảng 2 đưa ra sai số thống kê khi so sánh với kết quả tính toán từ DFT của các mô hình học máy trong dự đoán năng lượng của hệ vật liệu bcc-Fe ở trạng thái có từ tính (bao gồm cả tương tác hóa học và tương tác từ).

Kết quả từ bảng 2 cho thấy mô hình hồi quy tuyến tính cho kết quả tính năng lượng của tập huấn luyện rất tốt trong tất cả các phương pháp. Tuy nhiên với tập kiểm tra lại cho kết quả ngược lại là kém chính xác hơn các phương pháp còn lại ở hầu hết các chỉ số. Mô hình LASSO cho kết quả dự đoán tốt hơn ở tập kiểm tra nhưng lại kém hơn ở tập huấn luyện khi so với mô hình hồi quy Rigde. Nhìn chung, cả ba mô hình của chúng tôi đều dự đoán với độ chính xác rất cao năng lượng của các hệ vật liệu từ tính. Sự chênh lệch về các chỉ số thống kê đánh giá sai số là không nhiều. Một cách tổng thể, mô hình hồi quy Rigde cho kết quả phù hợp nhất.

Bảng 2. *Đánh giá sai số so với kết quả tính toán từ DFT của các mô hình dự đoán năng lượng của hệ ở trạng thái có từ tính*

Mô hình	Tập dữ liệu	MAE (eV/nguyên tử)	RMSE (eV/ nguyên tử)	R^2
Hồi quy tuyến tính	Huấn luyện	0,002261	0,040132	0,998918
	Kiểm tra	0,003004	0,043259	0,998484
Hồi quy LASSO	Huấn luyện	0,002753	0,045850	0,998683
	Kiểm tra	0,002562	0,044844	0,998707
Hồi quy Rigde	Huấn luyện	0,002313	0,041099	0,998893
	Kiểm tra	0,002710	0,042877	0,998633

4. Kết luận

Chúng tôi đã đề xuất mô hình, thực hiện các tính toán DFT để chuẩn bị dữ liệu và huấn luyện thành công các mô hình hồi quy (tuyến tính, Ridge và LASSO) có khả năng dự đoán nhanh năng lượng tương tác giữa các nguyên tử trong hệ vật liệu tinh thể bcc-Fe ở cả trạng thái có và không có từ tính. Các kết quả tính toán từ các mô hình học máy này đã được phân tích và so sánh với các kết quả tính toán bằng DFT, cho thấy độ chính xác cao. Trong đó, mô hình học máy sử dụng phương pháp hồi quy Ridge là tốt nhất. Có thể nói các mô hình học máy dựa trên hồi quy là một giải pháp tiềm năng giúp đơn giản hóa và rút ngắn thời gian cũng như chi phí tính toán năng lượng trong hệ vật liệu tinh thể bcc-Fe nói riêng và các hệ vật liệu từ nói chung.

Lời cảm ơn

Nghiên cứu này được tài trợ bởi trường Đại học Khoa học tự nhiên, Đại học Quốc gia Hà Nội trong đề tài mã số TN.23.06.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] P. Hohenberg and W. Kohn, "Inhomogeneous Electron Gas," *Phys. Rev.*, vol. 136, 1964, doi: 10.1103/PhysRev.136.B864.

[2] W. Kohn and L. J. Sham, "Self-Consistent Equations Including Exchange and Correlation Effects," *Phys. Rev.*, vol. 140, 1965, doi: 10.1103/PhysRev.140.A1133.

[3] C.M. Handley and J. Behler, "Next generation interatomic potentials for condensed systems," *Eur. Phys. J. B.*, vol. 87, 2014, Art. no. 152.

[4] A.P. Bartók and G. Csányi, "Gaussian approximation potentials: A brief tutorial introduction," *Int. J.*

Quantum Chem., vol. 115, pp. 1051-1057, 2015.
[5] S. De, A. P. Bartók, G. Csanyi, and M. Ceriotti, "Comparing molecules and solids across structural and alchemical space," *Phys. Chem. Chem. Phys.*, vol. 18, pp. 13754-13769, 2016.
[6] A. Seko *et al.*, "A sparse representation for potential energy surface," *Phys. Rev. B.*, vol. 90, 2014, Art. no. 24101.
[7] M. Rupp, A. Tkatchenko, K.-R. Muller, and O. A. von Lilienfeld, "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning," *Phys. Rev. Lett.*, vol. 108, 2012, Art. no. 58301.
[8] T. L. Pham, H. Kino, K. Terakura, T. Miyake, and H. C. Dam, "Novel mixture model for the representation of potential energy surfaces," *J. Chem. Phys.*, vol. 145, 2016, Art. no. 154103.
[9] J. Behler and M. Parrinello, "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces," *Phys. Rev. Lett.*, vol. 98, 2007, Art. no. 146401.
[10] N. Artrith and J. Behler, "High-dimensional neural network potentials for metal surfaces: A prototype study for copper," *Phys. Rev. B.*, vol. 85, 2012, Art. no. 45439.
[11] T.-C. Nguyen, V.-Q. Nguyen, V.-L. Ngoc, Q.-K. Than, and T.-L. Pham, "Learning hidden chemistry with deep neural networks," *Computational Materials Science*, vol. 200, 2021, Art. no. 110784.
[12] T. L. Pham, V. D. Nguyen, and T. C. Nguyen, "Machine Learning Representation for Atomic Forces and Energies," *VNU Journal of Science: Mathematics-Physics*, vol. 36, pp. 74-80, 2020.
[13] V.-Q. Nguyen, V.-C. Nguyen, T.-C. Nguyen, X.-V. Nguyen, and T.-L. Pham, "Pairwise interactions for potential energy surfaces and atomic forces using deep neural networks," *Computational Materials Science*, vol. 209, 2022, Art. no. 111379.
[14] Scikit-learn developers (BSD License), "Scikit-learn Machine Learning in Python". [Online]. Available: https://scikit-learn.org/stable/. [Accessed Jun. 25, 2024].
[15] S.-L. Shang, Y. Wang, and Z.-K. Liu, "Thermodynamic fluctuations between magnetic states from first-principles phonon calculations: The case of bcc Fe," *Phys. Rev. B*, vol. 82, 2010, Art. no. 014425.
[16] F. Kormann, A. Dick, B. Grabowski, T. Hickel, and J. Neugebauer, "Atomic forces at finite magnetic temperatures: Phonons in paramagnetic iron," *Phys. Rev. B*, vol. 85, 2012, Art. no. 125104.
[17] Y. Ikeda, A. Seko, A.Togo, and I. Tanaka, "Phonon softening in paramagnetic bcc Fe and its relationship to the pressure-induced phase transition," *Phys. Rev. B*, vol. 90, 2014, Art. no. 134106.