Лекция 11

Создание потенциалов межатомного взаимодействия на основе машинного обучения

Часть 1.

В данной части будет описан потенциал на основе машинного обучения, с помощью которого возможно с низкой ошибкой предсказывать энергию структур, создаваемых в ходе глобальной оптимизации в программе USPEX. Особенностью таких структур является то, что их энергии лежат в широком диапазоне, и на момент работы над таким потенциалом не существовало каких-либо других, способных решить эту задачу.

В данном потенциале энергия структуры раскладывается как сумма парной и многочастичной компонент:

$$E = E_0 + E_{2body} + E_{manybody},$$

В нашем подходе парный член восстанавливается линейной регрессией, а многочастичный – искусственной нейронной сетью (НС). Процесс обучения итеративный:

1 шаг – обучение только парного потенциала,

2 шаг – обучение многочастичной модели по разнице между реальной энергией и предсказанной парной частью,

3 шаг – обучение парной модели по разнице между реальной энергией и предсказанной многочастичной частью,

4 шаг – и так далее до сходимости.

Для того, чтобы оценить вклад парной части в общую энергию, для каждого потенциала мы рассчитывали следующий параметр:

$$\eta_{2body} = \langle \frac{E_{2body}}{E} \rangle$$

где среднее рассчитывалось по всем структурам. Если значение η близко к 1, значит система может быть описана с высокой точностью только парным потенциалом.

Выбор данных для обучения является одним из самых важных шагов в машинном обучении. Для того, чтобы обученный потенциал мог работать в режиме предсказания энергии в широком диапазоне значений, набор данных создавался с помощью программы USPEX.

На первом шаге мы создавали порядка 10000-20000 структур в одном поколении, энергии которых рассчитывались с высокой точностью программой VASP без ионной релаксации. Обучение на таких данных позволило получить хорошее первое приближение для весов линейной регрессии и нейронной сети. Следующий набор данных состоял из всех шагов релаксации для всех структур, созданных программой USPEX в ходе 10-20 поколений (30 структур на поколение, в итоге получалось около 30000 структур). При таком подходе финальный набор данных содержал как случайные, так и отрелаксированные структуры, что позволило обучить гибкий потенциал. Для того, чтобы алгоритм обучался только на различных структурах, мы применили простой классификатор из статьи [138]: если каждая компонента нового вектора признаков лежит в диапазоне значений, которые уже есть в обучающей выборке, то такой вектор признаков (а, соответственно, и структура) не добавляются в обучающую выборку.

Для парного и многочастичного вкладов в энергию рассматривались различные вектора признаков. В самом простом случае – парное взаимодействие может быть описано потенциалом Леннард-Джонса ($E = \epsilon\left[\frac{A}{r^{12}} - \frac{B}{r^6}\right]$). Мы же решили расширить этот подход и представить энергию парного взаимодействия как:

$$E_{2body} = \sum_{l=0}^{\infty} \sum_{i,j=1}^{N} \sum_{k=1}^{k_{max}} \frac{A_{i,j}^{k}}{r_{i,j}^{k}(l)}$$

где суммирование проводится внутри всего кристалла, l – индекс элементарной ячейки, i и j – индексы атомов в ячейке с $l = 0$ и текущим l, соответственно; k_{max} может принимать любые значения, мы в данном случае выбрали 15. Основная сложность заключается в том, что суммирование проводится по всему кристаллу и для $k \leq 6$ сумма сходится медленно. Для этого мы использовали метод Эвальда, описанный в статьях [150–152]. Таким образом, парная часть энергии является линейным произведением вектора коэффициентов $A^k_{i,j}$ на столбец $\frac{1}{r^k_{i,j}}$. Коэффициенты $A^k_{i,j}$ определяются из линейной регрессии. В качестве теста мы создали 10000 случайных структур типа A_xB_y в USPEX, энергия которых рассчитывалась программой GULP [152] с потенциалом Леннард-Джонса. У получившихся структур энергия лежала в диапазоне $[-30, -2]$ в единицах ϵ. Результаты восстановленных потенциалов для таких структур представлены на Рис. 3.1, RMSE составила $10^{-7}\epsilon$.

В многочастичную часть вектора признаков мы включили угловые функции из статьи [122]:

$$\sum_{j,k\neq i} \left(\frac{1+\lambda\cos\theta_{jik}}{2}\right)^{\xi} \exp(-\eta(R^2_{ij}+R^2_{ik}+R^2_{kj}))\cdot f_c(R_{ij})\cdot f_c(R_{ik})\cdot f_c(R_{kj}),$$

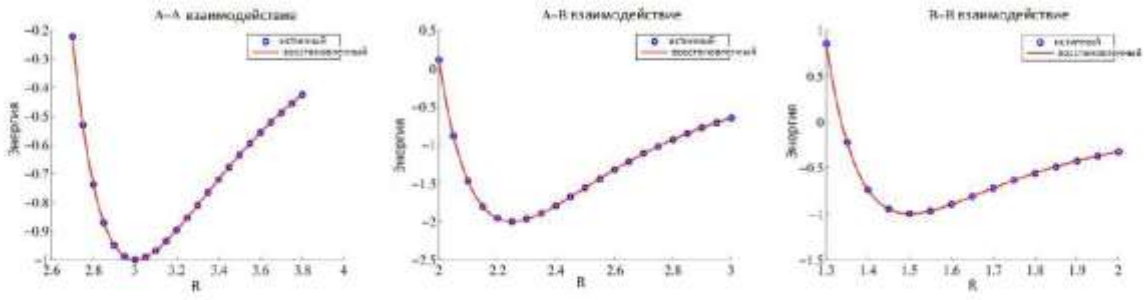


Рис. 3.1: Сравнение реальных и восстановленных потенциалов в случае системы A_xB_y с потенциалом Леннард-Джонса.

где суммирование происходит внутри сферы (с радиусом обрезания 6 Å), ϵ и η – свободные параметры, f_c – функция обрезания вида $0.5 \cdot [\cos(\frac{\pi x}{R_c}) + 1]$, если x меньше радиуса обрезания и 0, если больше. Также в многочастичную часть вектора признаков мы добавили объем элементарной ячейки и параметр порядка [153, 154]. В этих статьях также было показано, что параметр порядка коррелирует с энергией структуры: чем выше параметр порядка, тем ниже энергия. В сумме длина вектора признаков составляла 50.

Для восстановления многочастичной части энергии использовалась нейронная сеть с архитектурой (50-70-1). Мы тестировали и другие архитектуры, но эта показывала наилучшие результаты. В качестве функции активации использовалась функция $tanh(x) + \gamma x$. Веса нейронной сети обучались путем стандартного метода обратного распространения ошибки.

[11] Synthesis of FeH_5: A layered structure with atomic hydrogen slabs / CM Pépin, G Geneste, A Dewaele [и др.] // Science. 2017. Т. 357, № 6349. С. 382–385.

[12] Daw Murray S, Baskes Michael I. Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals // Phys. Rev. B. 1984. Т. 29, № 12. с. 6443.

[13] Behler J. Representing potential energy surfaces by high-dimensional neural network potentials // J. Phys. Condens. Matter. 2014. Т. 26, № 18. с. 183001.

[14] Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons / Albert P Bartók, Mike C Payne, Risi Kondor [и др.] // Phys. Rev. Lett. 2010. Т. 104, № 13. с. 136403.

[15] Shapeev Alexander V. Moment Tensor Potentials: a class of systematically improvable interatomic potentials // Multiscale Model. Simul. 2016. Т. 14, № 3. С. 1153–1173.

[16] Kresse Georg, Furthmüller Jürgen. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set // Phys. Rev. B. 1996. Т. 54, № 16. с. 11169.

[17] Kresse Georg, Hafner Jürgen. Ab Initio Molecular Dynamics for Liquid Metals // Phys. Rev. B. 1993. Т. 47, № 1. С. 558–561.

[18] Plimpton Steve. Fast Parallel Algorithms for Short-Range Molecular Dynamics // J. Comput. Phys. 1995. Т. 117. С. 1–19.

[19] Gavezzotti Angelo. Are crystal structures predictable? // Acc. Chem. Res. 1994. Т. 27, № 10. С. 309–314.

[20] Pickard Chris J., Needs R. J. High-Pressure Phases of Silane // Phys. Rev. Lett. 2006. Jul. T. 97. c. 045504. URL: https://link.aps.org/doi/10.1103/PhysRevLett.97.045504.

[21] Pickard Chris J, Needs R J. Ab initio random structure searching // J. Phys. Condens. Matter. 2011. T. 23, № 5. c. 053201. URL: http://stacks.iop.org/0953-8984/23/i=5/a=053201.

[22] Kirkpatrick S., Gelatt C. D., Vecchi M. P. Optimization by Simulated Annealing // Science. 1983. T. 220, № 4598. C. 671–680. URL: http://science.sciencemag.org/content/220/4598/671.

[23] Laio Alessandro, Parrinello Michele. Escaping free-energy minima // Proc. Natl. Acad. Sci. 2002. T. 99, № 20. C. 12562–12566. URL: http://www.pnas.org/content/99/20/12562.

[46] Tesauro Gerald. TD-Gammon, a self-teaching backgammon program, achieves master-level play // Neural Comput. 1994. T. 6, № 2. C. 215–219.

[115] Sinnott Susan B, Brenner Donald W. Three decades of many-body potentials in materials research // MRS Bull. 2012. T. 37, № 05. C. 469–473.

[116] Mishin Y, Lozovoi AY. Angular-dependent interatomic potential for tantalum // Acta Mater. 2006. T. 54, № 19. C. 5013–5026.

[117] Lipid14: the amber lipid force field / Callum J Dickson, Benjamin D Madej, Åge A Skjevik [и др.] // J. Chem. Theory Comput. 2014. T. 10, № 2. C. 865–879.

[118] CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields / Kenno Vanommeslaeghe, Elizabeth Hatcher, Chayan Acharya [и др.] // J. Comp. Chem. 2010. T. 31, № 4. C. 671–690.

[119] ReaxFF: a reactive force field for hydrocarbons / Adri CT Van Duin, Siddharth Dasgupta, Francois Lorant [и др.] // J. Phys. Chem. A. 2001. T. 105, № 41. C. 9396–9409.

[120] Lorenz Sönke, Groß Axel, Scheffler Matthias. Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks // Chem. Phys. Lett. 2004. T. 395, № 4. C. 210–215.

[121] Neural network models of potential energy surfaces / Thomas B Blank, Steven D Brown, August W Calhoun [и др.] // J. Chem. Phys. 1995. T. 103, № 10. C. 4129–4137.

[122] Behler Jörg, Parrinello Michele. Generalized neural-network representation of high-dimensional potential-energy surfaces // Phys. Rev. Lett. 2007. Т. 98, № 14. с. 146401.

[123] Ab initio quality neural-network potential for sodium / Hagai Eshet, Rustam Z Khaliullin, Thomas D Kühne [и др.] // Phys. Rev. B. 2010. Т. 81, № 18. с. 184107.

[124] Microscopic Origins of the Anomalous Melting Behavior of Sodium under High Pressure / Hagai Eshet, Rustam Z Khaliullin, Thomas D Kühne [и др.] // Phys. Rev. Lett. 2012. Т. 108, № 11. с. 115701.

[125] Metadynamics simulations of the high-pressure phases of silicon employing a high-dimensional neural network potential / Jörg Behler, Roman Martoňák, Davide Donadio [и др.] // Phys. Rev. Lett. 2008. Т. 100, № 18. с. 185501.

[126] Jose KV Jovan, Artrith Nongnuch, Behler Jörg. Construction of high-dimensional neural network potentials using environment-dependent atom pairs // J. Chem. Phys. 2012. Т. 136, № 19. с. 194111.

[127] Morawietz Tobias, Behler Jorg. A density-functional theory-based neural network potential for water clusters including van der Waals corrections // J. Phys. Chem. A. 2013. Т. 117, № 32. С. 7356–7366.

[128] Artrith Nongnuch, Morawietz Tobias, Behler Jörg. High-dimensional neural-network potentials for multicomponent systems: Application to zinc oxide // Phys. Rev. B. 2011. Т. 83, № 15. с. 153101.

[129] Szlachta Wojciech J, Bartók Albert P, Csányi Gábor. Accuracy and transferability of Gaussian approximation potential models for tungsten // Phys. Rev. B. 2014. Т. 90, № 10. с. 104108.

[130] Machine-learning approach for one-and two-body corrections to density functional theory: Applications to molecular and condensed water / Albert P Bartók, Michael J Gillan, Frederick R Manby [и др.] // Phys. Rev. B. 2013. Т. 88, № 5. с. 054104.

[131] Deringer Volker L., Csányi Gábor. Machine learning based interatomic potential for amorphous carbon // Phys. Rev. B. 2017. Mar. Т. 95. с. 094203. URL: https://link.aps.org/doi/10.1103/PhysRevB.95.094203.

[132] Deringer Volker L., Pickard Chris J., Csányi Gábor. Data-Driven Learning of Total and Local Energies in Elemental Boron // Phys. Rev. Lett. 2018. Apr. Т. 120. с. 156001. URL: https://link.aps.org/doi/10.1103/PhysRevLett.120.156001.

[133] Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials / Aidan P Thompson, Laura P Swiler, Christian R Trott [и др.] // J. Comp. Phys. 2015. Т. 285. С. 316–330.

[134] Podryabinkin Evgeny V, Shapeev Alexander V. Active learning of linear interatomic potentials // arXiv preprint arXiv:1611.09346. 2016.

[135] Gubaev Konstantin, Podryabinkin Evgeny V., Shapeev Alexander V. Machine learning of molecular properties: Locality and active learning // J. Chem. Phys. 2018. Т. 148, № 24. с. 241727.

[136] Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning / Evgeny V. Podryabinkin, Evgeny V. Tikhonov, Alexander V. Shapeev [и др.]. 2018.

[137] Big data meets quantum chemistry approximations: the Δ-machine learning approach / Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp [и др.] // J. Chem. Theory Comput. 2015. Т. 11, № 5. С. 2087–2096.

[138] Botu V., Ramprasad R. Ab-Initio Molecular Dynamics Acceleration Scheme with an Adaptive Machine Learning Framework // Int. J. Quantum Chem. 2014.

[139] Yao Kun, Herr John E, Parkhill John. The many-body expansion combined with neural networks // J. Chem. Phys. 2017. Т. 146, № 1. с. 014106.

[140] Behler Jörg. Perspective: Machine learning potentials for atomistic simulations // J. Chem. Phys. 2016. Т. 145, № 17. с. 170901.

[141] Mueller Tim, Kusne Aaron Gilad, Ramprasad Rampi. Machine learning in materials science: Recent progress and emerging applications // Rev. Comp. Ch. 2016. Т. 29. с. 186.

[142] Dickey J. M., Paskin A. Compute simulation of lattice dynamics of solids // Phys. Rev. 1969. Т. 188. С. 1407–1418.

[143] The self-consistent ab initio lattice dynamical method / Petros Souvatzis, Olle Eriksson, MI Katsnelson [и др.] // Comp. Mat. Sci. 2009. Т. 44, № 3. С. 888–894.

По материалам диссертации Круглова И.А. на тему «Поиск новых соединений, изучение их стабильности…»

Часть 2. Оценка производительности и стоимости машинного обучения межатомного потенциала

A Performance and Cost Assessment of Machine Learning Interatomic Potentials

Yunxing Zuo, Chi Chen, Xiangguo Li, Zhi Deng, Yiming Chen, and Shyue Ping Ong*

Department of NanoEngineering, University of California San Diego, 9500 Gilman Dr, Mail Code 0448, La Jolla, CA 92093-0448, United States

ВВЕДЕНИЕ

Началом для атомистического моделирования материалов является описание поверхности потенциальной энергии (ППЭ) как функции положения атомов. Хотя описания, основанные на квантовой механике, например, основанные на теории функционала плотности Кона-Шэма (DFT) [1,2], являются точными и применимыми для разных химикатов, их высокая стоимость и плохое масштабирование (обычно O (n3) или выше, где n - число электронов) [3–5] ограничивает моделирование ~ 1000 атомов и сотнями пикосекунд. Следовательно, крупномасштабные и длительные симуляции традиционно основываются на межатомных потенциалах (IAP), которые на сегодняшний день в большинстве случаев представляют собой эмпирические параметризации PES, основанные на физических функциональных формах, которые зависят только от атомных степеней свободы [6-8]. IAP получают линейное масштабирование по отношению к количеству атомов за счет точности и переносимости.

В последние годы появилась современная альтернатива в виде IAP с машинным обучением (ML-IAP), где PES описывается как функция дескрипторов локальной среды, инвариантных к трансляции, вращению и перестановке гомоядерных атомов [9, 10]. Примеры таких потенциалов включают:

-многомерный потенциал нейронной сети (NNP) [11, 12],

- потенциал гауссовского приближения (GAP) [10, 13, 14],

- потенциал спектрального анализа соседей (SNAP) [15–18],

- тензорные потенциалы моментов (MTP), [19–21],

- и другие [22–31].

Типичный подход к обучению таких потенциалов включает создание достаточно большого и разнообразного набора данных атомных конфигураций с соответствующими энергиями, силами и напряжениями, полученными из вычислений DFT, которые затем используются при обучении ML-IAP на основе одной или нескольких целей метрики, такие как минимизация средних абсолютных или квадратичных ошибок в прогнозируемых энергиях, силах, напряжениях или производных свойствах (например, упругих постоянных). Было показано, что ML-IAP являются значительным улучшением

по сравнению с традиционными IAP, в целом, обеспечивая точность, близкую к DFT, при прогнозировании энергий и сил для различных химических составов и атомных конфигураций. Тем не менее, критический пробел, который остается, - это строгая оценка относительных сильных и слабых сторон ML-IAP по стандартизированному набору данных, аналогично тому, что было сделано для классических IAP [32–34].

В этой работе мы представляем всестороннее сравнение производительности четырех основных ML-IAP.

- GAP, MTP, NNP и SNAP. Четыре IAP были оценены с точки зрения их точности в воспроизведении энергий и сил DFT, а также свойств материала, таких как уравнения состояния, параметры решетки и упругие постоянные. Также была предпринята попытка оценить требования к обучающим данным каждого ML-IAP и относительные вычислительные затраты на основе лучших доступных текущих реализаций.

Для обеспечения честного сравнения использовались стандартизированные наборы данных DFT для шести элементов (Li, Mo, Cu, Ni, Si и Ge) с одинаковыми обучающими / тестовыми выборками и аналогичными подходами к подгонке. Элементы были выбраны так, чтобы охватить различный химический состав и связи, например, металлы с ОЦК и ГЦК, металлы основной группы и переходные металлы, а также полупроводники группы IV.

II. МЕТОДЫ

A. Межатомные возможности машинного обучения

Четыре ML-IAP, исследованные в данной работе, уже подробно обсуждались в предыдущих работах и обзорах [9–21, 35–38]. Все ML-IAP выражают потенциальную энергию как сумму атомных энергий, которые являются функцией локальной среды вокруг каждого атома, но различаются дескрипторами для этих локальных сред и подходом / функциональным выражением ML, используемым для сопоставления дескрипторов с потенциальным энергия. Подробный формализм всех четырех ML-IAP представлен в Дополнительной информации. Здесь только краткое изложение ключевых концепций и параметров модели, лежащих в основе ML-IAP в хронологическом порядке развития, предоставляется, чтобы помочь читателю следовать оставшейся части этой статьи.

1. Многомерный нейросетевой потенциал (NNP). NNP использует атомно-центрированные функции симметрии (ACSF) [39] для представления атомной локальной среды и полносвязные нейронные сети для описания PES относительно функций симметрии [11, 12]. Для каждого атома используется отдельная нейронная сеть. Нейронная сеть определяется количеством скрытых слоев и узлов в каждом слое, в то время как пространство дескрипторов задается следующими функциями симметрии:

$$G_i^{\text{atom,rad}} = \sum_{j \neq i}^{N_{\text{atom}}} e^{-\eta(R_{ij}-R_s)^2} \cdot f_c(R_{ij}), \tag{1}$$

$$G_i^{\text{atom,ang}} = 2^{1-\zeta} \sum_{j,k \neq i}^{N_{\text{atom}}} (1 + \lambda \cos\theta_{ijk})^{\zeta} \cdot e^{-\eta'(R_{ij}^2+R_{ik}^2+R_{jk}^2)} \cdot f_c(R_{ij}) \cdot f_c(R_{ik}) \cdot f_c(R_{jk}) \tag{2}$$

где Rij - расстояние между атомом i и соседним атомом j, η - ширина гауссиана, а Rs - сдвиг положения по всем соседним атомам в пределах границы радиус Rc, ηJ - ширина гауссова базиса, а ζ - угловое разрешение.

fc (Rij) - функция отсечения, определяемая следующим образом:

$$f_c(R_{ij}) = \begin{cases} 0.5 \cdot [\cos(\frac{\pi R_{ij}}{R_c}) + 1], & \text{for } R_{ij} \leq R_c \\ 0.0, & \text{for } R_{ij} > R_c. \end{cases} \tag{3}$$

Эти гиперпараметры были оптимизированы, чтобы минимизировать средние абсолютные ошибки энергий и сил для каждого химического режима. Модель NNP показала отличные характеристики для Si [11], TiO2 [40], воды [41] и границ раздела твердое тело-жидкость [42], металл-органическое соединение[43], и был расширен, чтобы включить электростатику дальнего действия для ионных систем, таких как ZnO [44] и Li3PO4 [45].

2. Потенциал гауссовой аппроксимации (GAP). GAP вычисляет сходство между атомными конфигурациями на основе ядра с плавным перекрытием атомных позиций (SOAP) [10, 46], которое затем используется в гауссовой модели процесса. В SOAP плотности атомных соседей ρi (R), размазанные по Гауссу, разлагаются в сферические гармоники следующим образом:

$$\rho_i(\boldsymbol{R}) = \sum_j f_c(R_{ij}) \cdot \exp(-\frac{|\boldsymbol{R} - \boldsymbol{R}_{ij}|^2}{2\sigma_{\text{atom}}^2}) = \sum_{nlm} c_{nlm}\ g_n(R) Y_{lm}(\hat{\boldsymbol{R}}), \tag{4}$$

Вектор сферического спектра мощности, который, в свою очередь, является квадратом коэффициентов разложения,

$$p_{n_1 n_2 l}(\boldsymbol{R}_i) = \sum_{m=-l}^{l} c_{n_1 lm}^* c_{n_2 lm}, \tag{5}$$

может использоваться для построения ядра SOAP при возведении в положительную целую степень ζ(который в данном случае равен 4), чтобы подчеркнуть чувствительность ядра [10],

$$\boldsymbol{K}(\boldsymbol{R},\boldsymbol{R'}) = \sum_{n_1 n_2 l} (p_{n_1 n_2 l}(\boldsymbol{R}) p_{n_1 n_2 l}(\boldsymbol{R'}))^{\zeta}, \qquad (6)$$

В приведенных выше уравнениях σatom - это гладкость, контролирующая гауссово размытие, и nmax и lmax определяют максимальные мощности для радиальных компонентов и угловой компоненты.компоненты в разложении по сферическим гармоникам соответственно [10].Эти гиперпараметры, а также количество эталонных атомных , используемых в гауссовском процессе, оптимизированы в процедуре настройки для получения оптимальной производительности. ГАП был разработан для переходных металлов [13, 14], элементов основных групп [47–49], алмазных полупроводников [50, 51], а также для многокомпонентных систем [37].

3. Возможности спектрального анализа соседей (SNAP). SNAP использует коэффициенты биспектра функций плотности атомных соседей [10] в качестве дескрипторов. В исходной формулировке SNAP предполагается линейная модель между энергиями и компонентами биспектра [15]. Недавно была разработана квадратичная модель (обозначаемая в этой работе как qSNAP) [52], которая расширяет линейную модель энергии SNAP, чтобы включить все различные попарные произведения компонентов биспектра. В этой работе были исследованы как линейные, так и квадратичные модели SNAP. Ключевыми гиперпараметрами, влияющими на производительность модели, являются радиус обрезания и Jmax, ограничивающий индексы j1, j2,j в коэффициентах связи Клебша-Гордана H jmmj2m2mJ2 в построении биспектра составные части:

$$B_{j_1,j_2,j} = \sum_{m_1,m_1'=-j_1}^{j_1} \sum_{m_2,m_2'=-j_2}^{j_2} \sum_{m,m'=-j}^{j} (u^{j}_{m,m'})^{*} \times H^{jmm'}_{j_1 m_1 m_1' \, j_2 m_2 m_2'} u^{j_1}_{m_1,m_1'} u^{j_2}_{m_2,m_2'}, \qquad (7)$$

где uj являются коэффициентами в 4-мерном разложении по гиперсферическим гармоникам функция плотности соседей:

$$\rho_i(\mathbf{R}) = \sum_{j=0}^{\infty} \sum_{m,m'=-j}^{j} u^{j}_{m,m'} U^{j}_{m,m'}, \qquad (8)$$

Модель SNAP, а также модель qSNAP продемонстрировали большой успех в переходных металлах [15–17, 52], а также в бинарных системах [17, 18, 38].

4. Тензорный потенциал момента (MTP). MTP [19] разрабатывает вращательно-ковариантные тензоры

$$M_{\mu,\nu}(\boldsymbol{R}) = \sum_j f_\mu(R_{ij}) \underbrace{\boldsymbol{R}_{ij} \otimes \cdots \otimes \boldsymbol{R}_{ij}}_{\nu \ times}, \qquad (9)$$

для описания атомных локальных сред. Здесь fμ - радиальные функции, а Rij

а Rij ⊗ • • • ⊗ Rij - тензоры ранга ν, кодирующие угловую информацию об атомном окружении. Ранг ν может быть достаточно большим, чтобы аппроксимировать любые произвольные взаимодействия. Затем MTP сжимает эти тензоры до скаляра, получая вращательно-инвариантный базис функций и применяет линейную регрессию для корреляции энергий с базисными функциями. Производительность MTP контролируется полиномиальной степенной метрикой, которая определяет, какие тензоры и сколько раз сжимаются. Модель MTP успешно применялась к металлам [19, 20, 53], бору [54], Б. Наборы данных DFT

Полный набор данных DFT был создан для шести элементов - Li, Mo, Ni, Cu, Si и Ge. Эти элементы были выбраны, чтобы охватить множество химических элементов (металл основной группы, переходный металл и полупроводник), кристаллических структур (ОЦК, ГЦК и алмаз) и типов связи (металлических и ковалентных). Для каждого элемента мы сгенерировали набор структур с различным охватом атомарного пространства локальной среды, а именно:

(1) Кристалл в основном состоянии для каждого элемента.

(2) Напряженные структуры, построенные путем приложения деформаций от −10% до 10% с интервалами 2% к объемной сверхъячейке в шести различных режимах, как описано в работе de Jong et al. [56]. Используемые суперячейки представляют собой 3 × 3 × 3, 3 × 3 × 3 и 2 × 2 × 2 обычных элементарных ячеек с ОЦК, ГЦК и алмазами соответственно.

(3) Плиты с максимальным индексом Миллера, равным трем, включая (100), (110), (111), (210), (211), (310), (311), (320), (321) , (322), (331) и (332), полученные из базы данных Crystalium [57, 58].

(4) NVT ab initio моделирование молекулярной динамики (AIMD) объемных сверхъячеек (аналогично (2)), выполненное при 300 K и 0,5 ×, 0,9 ×, 1,5 ×, 2,0 × точки плавления каждого элемента. В общей сложности 20 снимков были получены из каждого моделирования AIMD с интервалом 0,1 пс, если не указано иное.

(5) NVT AIMD-моделирование объемных сверхъячеек (аналогично (2)) с единственной вакансией, выполненное при 300 K и 2,0-кратной температуре плавления каждого элемента. В общей сложности 40 снимков были получены из каждого моделирования AIMD с интервалом 0,1 пс, если не указано иное.

Все расчеты DFT были выполнены с использованием пакета программ Венского ab initio моделирования (VASP) [59] версии 5.4.1 в рамках подхода с расширенной волной проектора [60]. Для обменно-корреляционного функционала было принято приближение обобщенного градиента (GGA) Perdew-Burke-Ernzerhof (PBE) [61]. Ограничение кинетической энергии было установлено на 520 эВ, а k-точка

Размер сетки составлял 4 × 4 × 4 для суперячейек из Mo, Ni, Cu, Si и Ge и 3 × 3 × 3 для суперячейек из Li. Составляющие электронной энергии и атомной силы сведены с точностью до 10-5 эВ и

0,02 эВ / А соответственно, согласно предыдущим работам [16, 17]. Моделирование AIMD было

выполнялись с одной точкой Γ k и не были спин-поляризованными, но статические расчеты с использованием тех же параметров, что и остальные данные, проводились на снимках для получения согласованных энергий и сил. Все структурные манипуляции и анализ вычислений DFT проводились с использованием библиотеки Python Materials Genomics (Pymatgen) [62], а автоматизация вычислений выполнялась с помощью программного обеспечения Fireworks [63].двойным и тройным сплавам [21], а также к химическим реакциям в газовой фазе [55].

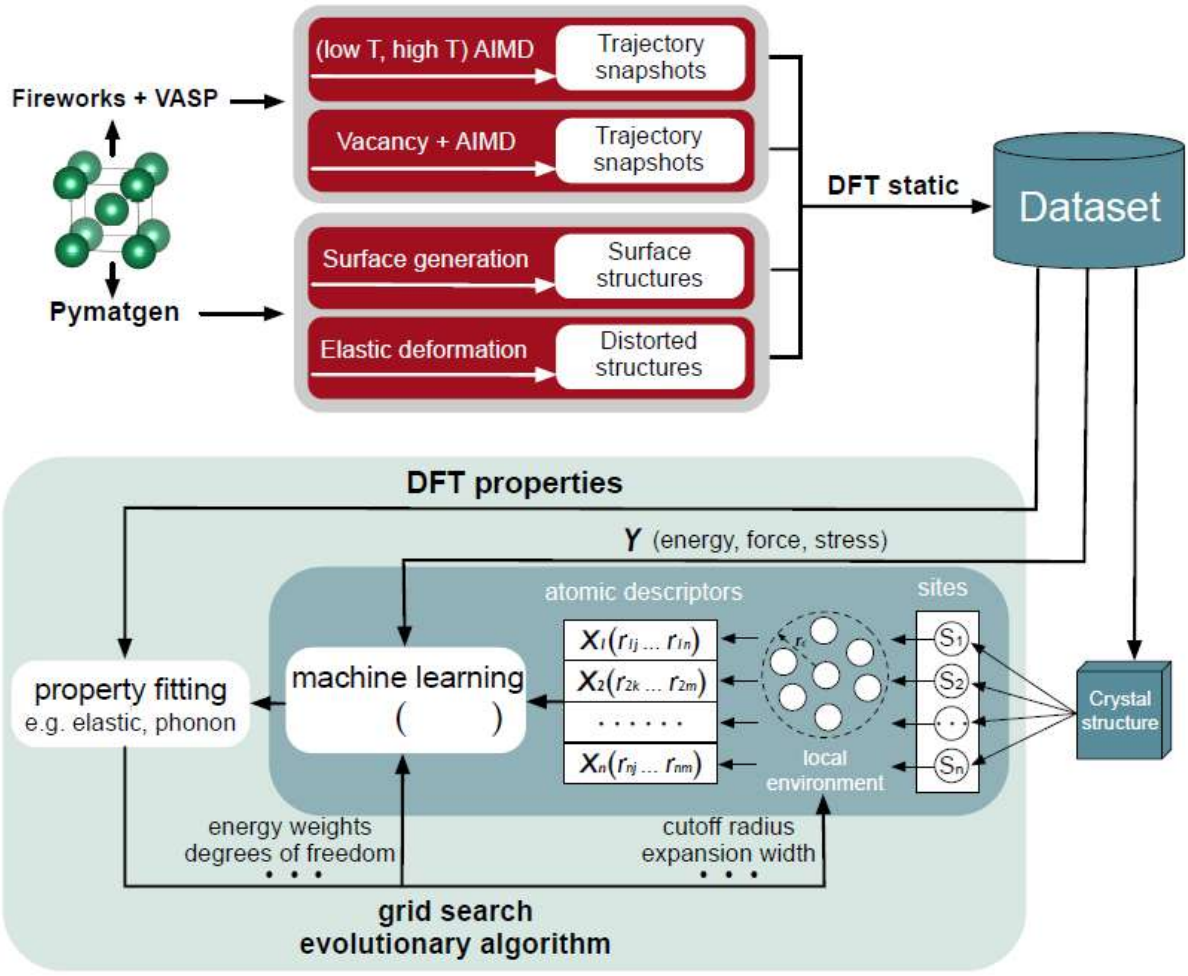


РИС. 1. Рабочий процесс разработки межатомного потенциала машинного обучения.

На рисунке 1 представлен обзор генерации общих данных и потенциального развития.

схема. Набор обучающих данных был сначала создан посредством статических вычислений DFT для четырех категорий структур. Процедура оптимизации состояла из двух циклов. Во внутреннем цикле выбранные структуры в базе данных были преобразованы в атомарные дескрипторы (например, компоненты биспектра для SNAP и функции симметрии для NNP), которые затем были введены в соответствующую модель ML вместе с энергиями, силами и напряжениями DFT в качестве цели обучения. Данные были распределены по обучающей и тестовой выборкам с разделением 90:10. Параметры моделей ML были оптимизированы в процессе обучения. Во внешнем цикле модель ML, обученная во внутреннем цикле, использовалась для прогнозирования основных свойств материала (например, тензоров упругости), а затем различия между прогнозируемыми и эталонными значениями использовались для определения оптимальных гиперпараметров для каждого ML-IAP. В этой работе мы

приняла комбинацию алгоритма поиска по сетке и алгоритма дифференциальной эволюции для выполнения оптимизации гиперпараметров для различных ML-IAP.

D. Доступность данных и кода

Чтобы облегчить повторное использование и воспроизведение наших результатов, код, данные и оптимизированные модели машинного обучения в этой работе опубликованы с открытым исходным кодом на Github (https://github.com/ materialsvirtuallab / mlearn). Код включает высокоуровневые интерфейсы Python для разработки ML-IAP, а также калькуляторы свойств материалов LAMMPS.

III. ПОЛУЧЕННЫЕ РЕЗУЛЬТАТЫ

A. Оптимизированные параметры модели

Оптимизированные коэффициенты и гиперпараметры для каждого ML-IAP приведены в дополнительной информации (см. Таблицу S1 – Таблицу S10). Здесь мы ограничим наши обсуждения параметром, общим для всех ML-IAP - радиусом отсечения - и представим исследование сходимости каждого ML-IAP с количеством степеней свободы модели.

Радиус отсечки определяет максимальный диапазон межатомных взаимодействий и, следовательно, имеет решающее влияние на характеристики предсказания ML-IAP. В таблице I представлены оптимизированные радиусы отсечки различных ML-IAP для разных химикатов. Различные ML-IAP дают одинаковые оптимизированные радиусы отсечки для одной и той же элементной системы. Оптимизированные радиусы отсечки находятся между вторым ближайшим соседом (2NN) и расстоянием 3NN для элементов с ГЦК (Cu, Ni), между расстояниями 3NN и 4NN для элементов с ОЦК (Li, Mo) и алмаза (Ge и Si). Эти наблюдения согласуются с предыдущими исследованиями традиционных и ML IAP, где обычно 2NN-взаимодействия оказываются достаточными для ГЦК-металлов [64, 65], в то время как вклады 3NN нельзя игнорировать для ОЦК-металлов [13, 19, 20 , 66, 67] и алмазных систем [68, 69].

Количество степеней свободы (DOF), например, количество весов и смещений для NNP и количество репрезентативных точек в GAP, сильно влияет на точность и вычислительные затраты каждого ML-IAP. На рисунке 2 показан компромисс между вычислительными

	fcc		bcc		diamond	
cutoff radius (Å)	Ni	Cu	Li	Mo	Si	Ge
GAP	3.9	3.9	4.8	5.2	5.4	5.4
MTP	4.0	3.9	5.1	5.2	4.7	5.1
NNP	3.9	4.1	5.2	5.2	5.2	5.6
SNAP	3.9	4.1	5.1	4.6	4.9	5.5
qSNAP	3.8	3.9	5.1	5.2	4.8	4.9

TABLE I. Optimized cutoff radius for each element for each ML-IAP.

АБЛИЦА I. Оптимизированный радиус отсечки для каждого элемента для каждого ML-IAP.

стоимость и ошибка тестирования при различных степенях свободы для каждого установленного Mo ML-IAP. Аналогичные результаты получены для других систем (см. Рисунок S7 – Рисунок S11). Следует отметить, что относительные вычислительные затраты основаны на наиболее эффективных доступных реализациях [10, 15, 19, 52, 70] каждого ML-IAP в настоящее время в LAMMPS [71] и выполняются на одном ядре ЦП Intel. i7-6850k 3,6 ГГц с объемной сверхъячейкой 18 × 18 × 18, содержащей 23 328 атомов для системы Mo. Будущие реализации могут улучшить эти результаты. Граница Парето проведена на рисунке 2a для представления точек, в которых лучшая точность может быть достигнута только ценой более высоких вычислительных затрат [72], а черные стрелки указывают «оптимальные» конфигурации для каждой модели с точки зрения компромисса между ошибка теста и вычислительные затраты. Эти «оптимальные» конфигурации использовались для последующего сравнения точности энергий, сил и свойств. Мы обнаружили, что «оптимальные» модели MTP, NNP, SNAP и qSNAP имеют тенденцию быть на два порядка меньше в вычислительном отношении, чем «оптимальная» модель GAP. Модели MTP обычно лежат близко к границе Парето, демонстрируя отличный баланс между точностью модели и вычислительной эффективностью. Для моделей SNAP и qSNAP пространство дескрипторов (то есть биспектральные компоненты) определяется параметром Jmax. Мы обнаружили, что ограничивающим скорость шагом является вычисление биспектра, а вычисление квадратичных членов в qSNAP оказывает лишь небольшое влияние на вычислительные затраты [52]. Однако мы обнаруживаем, что существенное расширение числа подгоняемых коэффициентов в модели qSNAP приводит к большей вероятности переобучения, особенно для Jmax> 3 (см. Рисунок 2b). Для модели GAP вычислительные затраты линейно связаны с количеством ядер, используемых в регрессии гауссовского процесса [13].

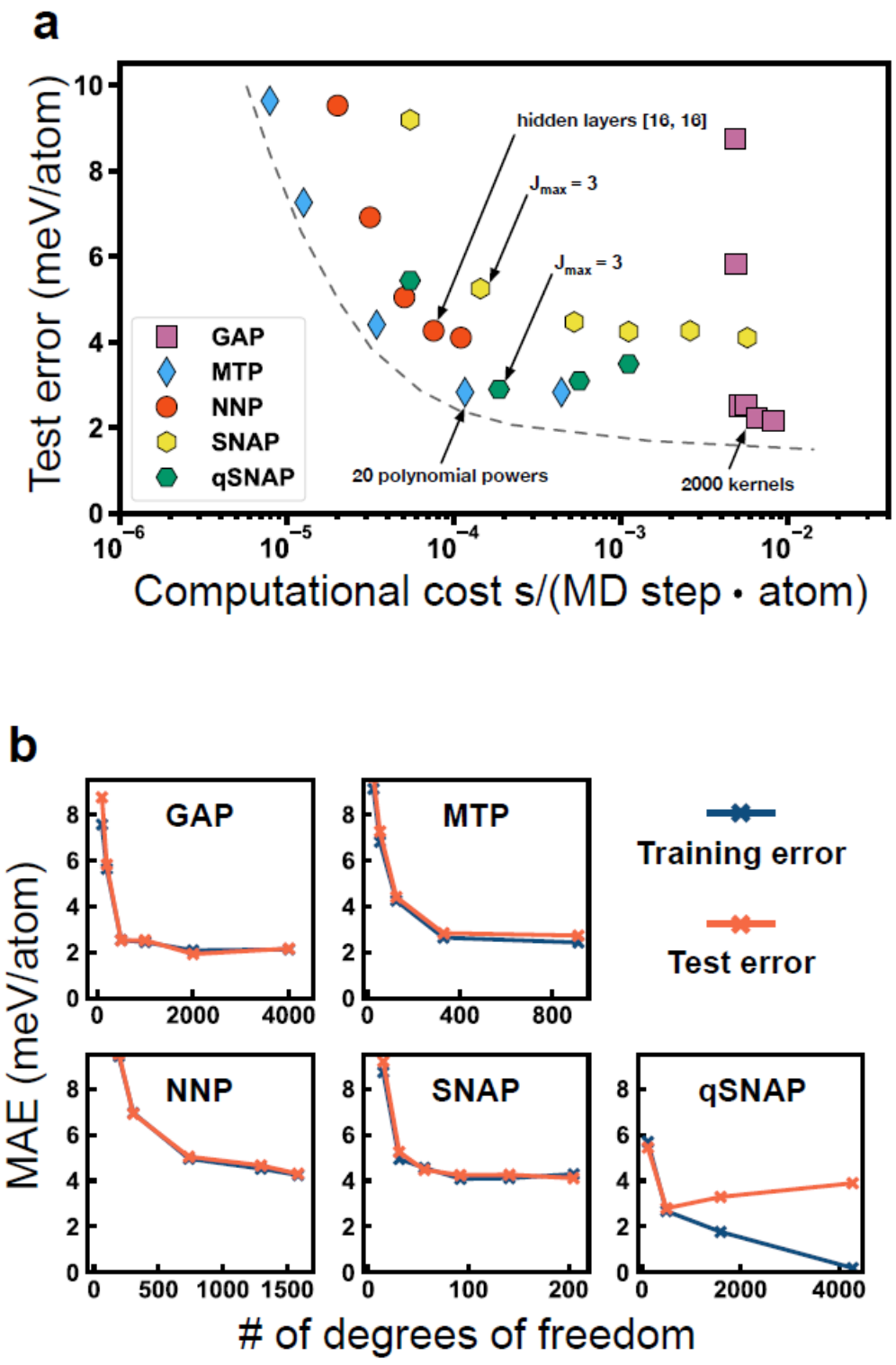


РИС. 2. (a) Ошибка теста в сравнении с вычислительными затратами для системы Mo.

Черная пунктирная линия обозначает Парето.

граница, представляющая оптимальный компромисс между точностью и вычислительными затратами. Сроки были выполняется расчетами LAMMPS на одном ядре процессора Intel i7-683,6 ГГц. Черные стрелки обозначают «оптимальную»

конфигурацию для каждого ML-IAP, который использовался в последующих сравнения. (б) Графики ошибок обучения и тестирования в зависимости от количества степеней свободы для каждого МЛ-ИАП.

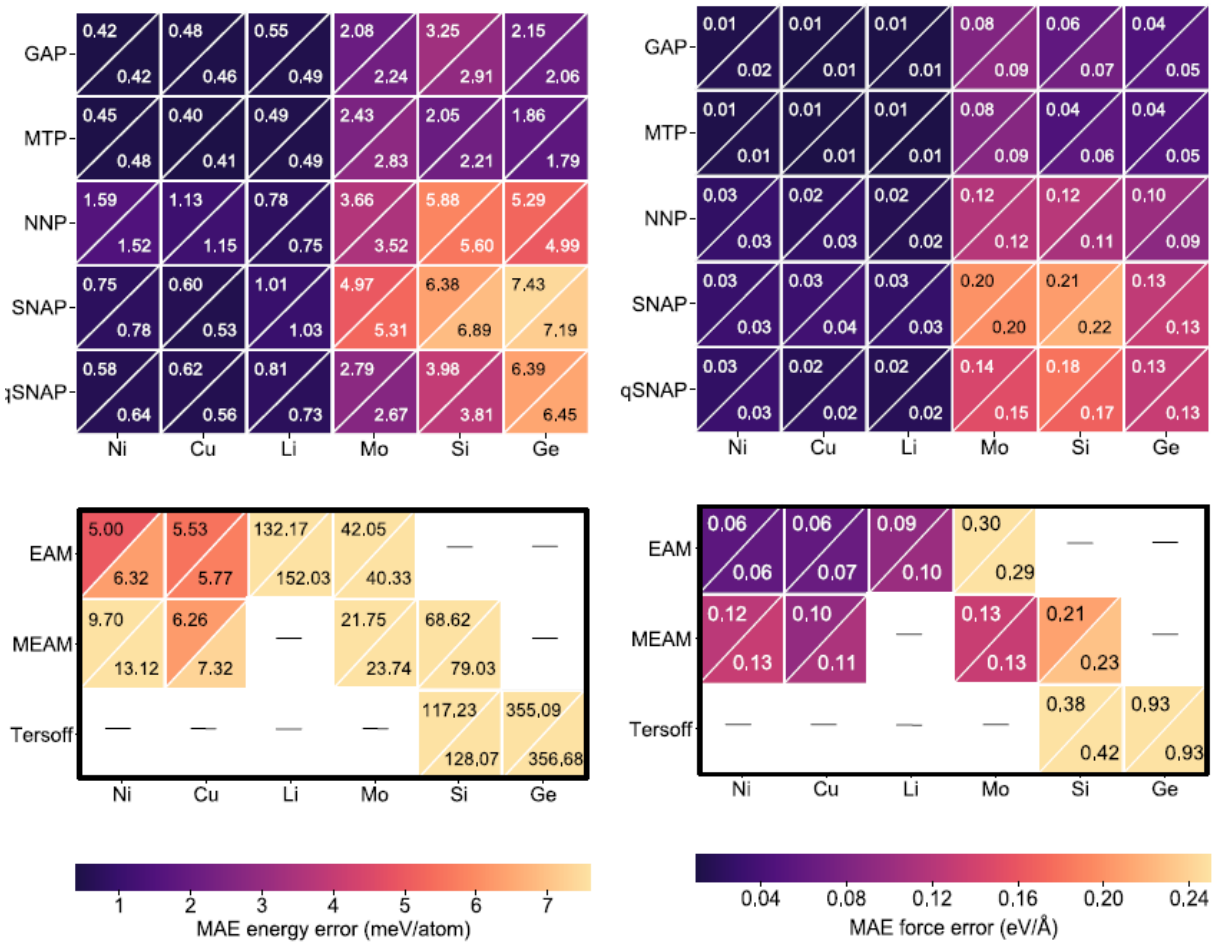
Б. Точность в энергиях и силах

На рис. 3 представлено сравнение MAE в энергии и сил для четырех ML-IAP и лучших доступных классических IAP относительно DFT. Все ML-IAP демонстрируют чрезвычайно хорошую производительность по всем изученным элементам, достигая MAE по энергии и силам, которые намного ниже, чем у лучших доступных традиционных IAP для каждого элемента. Следует отметить, что различия в МАЭ между ML-IAP находятся в масштабе мэВ атом − 1 по энергии и 0,1 эВ А − 1 по силе; следовательно, любое последующее обсуждение относительных характеристик ML-IAP следует рассматривать в контексте того, что даже самые большие различия в точности между ML-IAP уже близки к пределам ошибки DFT. Во всех случаях ошибки обучения и тестирования схожи, что указывает на отсутствие чрезмерной подгонки для оптимизированных ML-IAP.

Модели GAP и MTP обычно имеют самые низкие значения MAE по энергиям и силам. Наибольшие значения МАЭ по энергиям наблюдаются для моделей SNAP и NNP. Хорошо известно, что модели на основе нейронных сетей часто требуют больших наборов данных для лучшей производительности; предыдущие модели NNP были обучены на тысячах или десятках тысяч структур [73, 74], в то время как только сотни структур используются для обучения текущих ML-IAP. Тем не менее, модели NNP по-прежнему показывают удивительно хорошую производительность для систем bcc. Производительность моделей qSNAP находится между GAP и NNP. В целом, модели qSNAP имеют умеренно более низкие значения MAE, чем линейный SNAP, хотя и за счет значительного увеличения числа параметров.

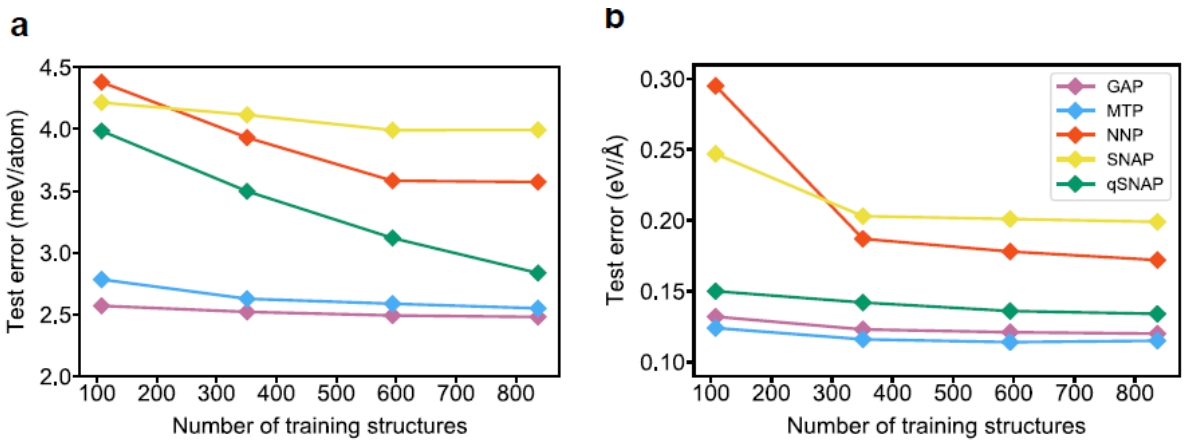
С точки зрения химии, мы обнаружили, что самые низкие значения МАЭ по энергии наблюдаются для систем с ГЦК, за ними следуют системы с ОЦК, а самые высокие значения МАЭ наблюдаются для алмазных систем. Очень низкие значения MAE в силах наблюдаются для всех ML IAP для Cu, Ni и Li, в то время как значительно более высокие MAE в силах наблюдаются для Mo, металла с более высоким модулем и большим распределением сил. Более высокие силы МАЭ наблюдаются и для алмазных полупроводников. Эти тенденции в целом одинаковы для всех изученных ML-IAP.

Мы также провели исследование сходимости ML-IAP с размером обучающих данных, используя Mo в качестве эталонной системы, учитывая, что это металл с ОЦК (для которого традиционные IAP, как правило, плохо работают) с большим распределением сил. Здесь продолжительность моделирования AIMD была увеличена в четыре раза, и за тот же интервал времени было выбрано больше обучающих структур. Результаты сходимости показаны на рисунке 4. Хотя ошибки прогнозирования всех моделей уменьшаются с увеличением количества обучающих структур, наиболее существенные



(a) (a) Средние абсолютные ошибки в прогнозируемых энергиях (б) Средние абсолютные ошибки в прогнозируемых силах

РИС. 3. Средние абсолютные ошибки в (а) прогнозируемых энергиях (б) прогнозируемых сил для всех четырех ML-IAP, а также традиционные IAP (EAM [75, 76], MEAM [77–79], Tersoff [80, 81]). Верхний левый и нижний правый треугольники в каждой ячейке представляют собой ошибки обучения и тестирования соответственно. Повышение точности, особенно в прогнозируемых энергиях, наблюдается для моделей NNP и qSNAP. Модель SNAP Mo, похоже, сошлась по энергии и силе при размере обучающих данных ~ 600 и ~ 400 структур соответственно. Для NNP дополнительные тренировочные структуры предлагают скромные улучшения в точности силы, но большие улучшения в точности энергии. Действительно, возможно, что модели NNP и qSNAP Mo не были сведены в отношении точности по энергиям даже при ~ 800 обучающих структурах. Мы не пытались еще больше сблизить эти модели ввиду связанных с этим вычислительных затрат.



. Точность свойств материала

Точность прогнозирования основных свойств материала имеет решающее значение для оценки эффективности ML-IAP. Здесь мы выполняем метод подталкиваемой эластичной ленты с восходящим изображением (CI-NEB) [82], а также молекулярную динамику (MD) с ML-IAP, чтобы получить параметр кубической решетки, упругие константы, энергии миграции и энергии образования вакансий. Сравнение этих прогнозируемых свойств материала со значениями DFT представлено в Таблице II. Характеристики всех ML-IAP в целом превосходны: параметры решетки находятся в пределах 0,1–2,0% от значений DFT, а упругие константы обычно находятся в пределах 10% от значений DFT. Следует отметить, что большая процентная погрешность Li для упругих постоянных связана с небольшими контрольными значениями. Модели MTP, SNAP и qSNAP хорошо работают с упругими константами в системах с ГЦК и ОЦК, но демонстрируют несколько более высокие ошибки прогноза в алмазных системах. Возможным объяснением несколько худшего предсказания упругих констант модели NNP может быть ограничение размера обучающих данных, которые ограничивают потенциал полносвязной нейронной сети. Однако следует отметить, что, несмотря на несколько более высокие ошибки предсказания упругих компонентов для модели NNP, ее ошибки предсказания аппроксимированного модуля упругости Фойгта-Рейсс-Хилла [83] для различных элементарных систем хорошо согласуются с эталонными значениями DFT.

Что касается диффузионных свойств, модели GAP и MTP хорошо работают в разных химических структурах, при этом большинство ошибок прогноза находятся в пределах 10% от значений DFT, хотя умеренная недооценка энергии миграции для алмазных систем в соответствии с предыдущим исследованием [50]. Хотя модели SNAP и qSNAP показывают высокую точность в предсказании диффузионных свойств для систем с ГЦК, они значительно недооценивают энергию образования вакансий, а также активационный барьер для алмазных систем. Примечательно, что все ML-IAP переоценивают энергию миграции системы Mo более чем на 20%, что также наблюдалось в предыдущей работе [16].

ТАБЛИЦА II: Расчетный параметр кубической решетки a, упругие постоянные (cij), объемный модуль Фойгта-Рейсс-Хилла

BVRH, энергия миграции (Em), энергия образования вакансии (Em), а также активационный барьер для вакансии

диффузия (Ea = Ev + Em) с DFT и четырьмя ML-IAP. Наименьшие абсолютные ошибки относительно

для каждого свойства выделены жирным шрифтом для удобства использования. Проценты ошибок относительно значений DFT равны

показаны в скобках.

	DFT	GAP	MTP	NNP	SNAP	qSNAP
Ni						
a (Å)	3.508	3.523 (0.4%)	3.522 (0.4%)	3.523 (0.4%)	3.522 (0.4%)	**3.521 (0.4%)**
c_{11} (GPa)	276	281 (1.8%)	284 (2.9%)	**274 (-0.8%)**	283 (2.5%)	267 (−3.3%)
c_{12} (GPa)	159	**159 (0.0%)**	172 (8.2%)	169 (6.3%)	168 (5.7%)	155 (−2.5%)
c_{44} (GPa)	132	126 (−4.5%)	127 (-3.8%)	113 (−14.4%)	**129 (-2.3%)**	125 (−5.3%)
B_{VRH} (GPa)	198	**200 (1.0%)**	209 (5.6%)	204 (3.0%)	206 (4.0%)	193 (−2.5%)
E_v (eV)	1.49	1.46 (−2.0%)	1.43 (−4.0%)	1.65 (10.7%)	**1.47 (-1.3%)**	**1.47 (-1.3%)**
E_m (eV)	1.12	1.14 (1.8%)	1.11 (−0.9%)	1.14 (1.8%)	**1.12 (0.0%)**	1.05 (−6.3%)
E_a (eV)	2.61	**2.60 (-0.4%)**	2.54 (−2.7%)	2.79 (6.9%)	2.59 (−0.8%)	2.52 (−3.4%)
Cu						
a (Å)	3.621	**3.634 (0.4%)**	3.636 (0.4%)	3.637 (0.4%)	3.634 (0.4%)	**3.636 (0.4%)**
c_{11} (GPa)	173	**175 (1.2%)**	177 (2.3%)	182 (5.2%)	178 (2.9%)	**178 (2.9%)**
c_{12} (GPa)	133	120 (−9.8%)	120 (9.8%)	125 (−6.0%)	126 (−5.3%)	124 (−6.8%)
c_{44} (GPa)	88	82 (−6.8%)	81 (−8.0%)	76 (−13.6%)	**86 (-2.3%)**	82 (−6.8%)
B_{VRH} (GPa)	146	138 (−5.5%)	139 (−4.8%)	144 (−1.4%)	**143 (-2.1%)**	142 (−2.7%)
E_v (eV)	1.15	1.05 (−8.7%)	1.10 (−4.3%)	1.23 (7.0%)	1.19 (3.5%)	**1.15 (0.0%)**
E_m (eV)	0.79	0.76 (−3.8%)	**0.77 (-2.5%)**	**0.77 (-2.5%)**	0.82 (3.8%)	0.74 (−6.3%)
E_a (eV)	1.94	1.81 (−6.7%)	1.87 (−3.6%)	2.00 (3.1%)	2.01 (3.6%)	**1.89 (-2.6%)**

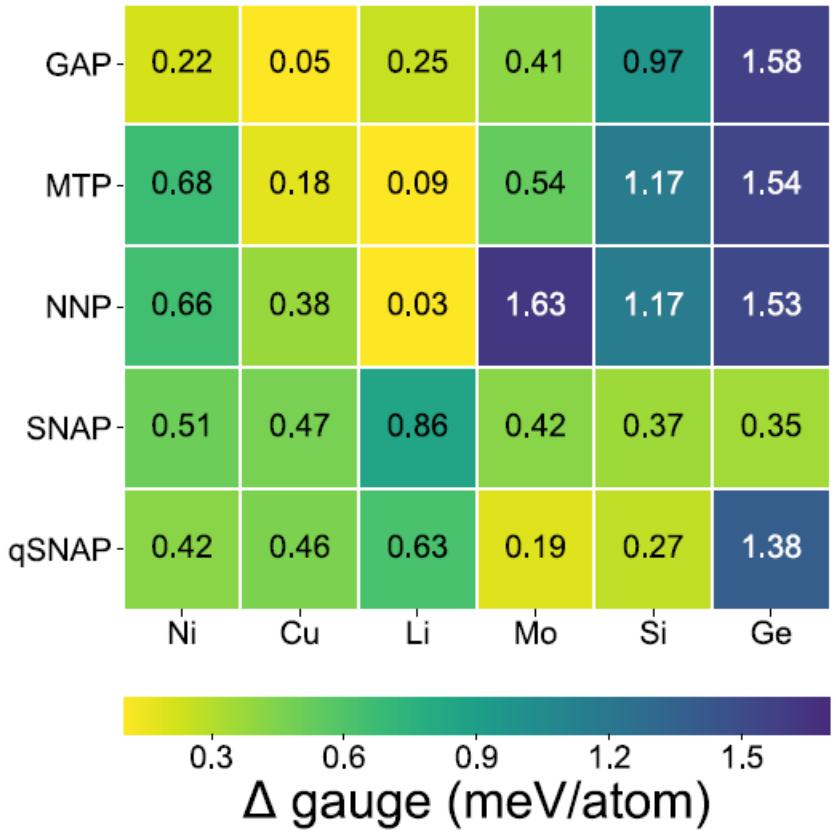
D. Точность в уравнениях состояния

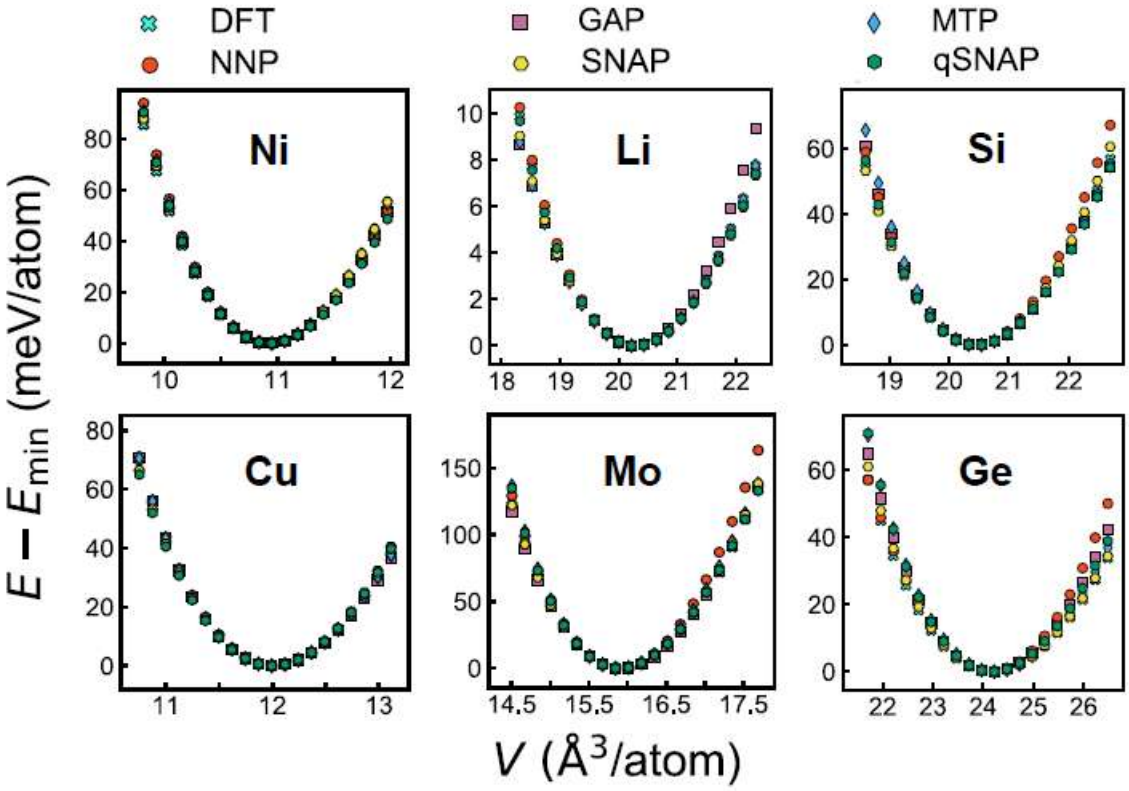
Чтобы обеспечить оценку эффективности ML-IAP вдали от равновесия, мы вычислили попарное сравнение кривых уравнения состояния (EOS) для всех изученных элементов с использованием датчика ΔEOS Lejaeghere et al. [84–86] Датчик ΔEOS, который использовался для оценки разницы в точности между кодами DFT, представляет собой среднеквадратическую разницу между двумя кривыми EOS в интервале ± 6% от равновесного объема, определяемую следующим образом:

$$\Delta_{\mathrm{EOS}} = \sqrt{\frac{\int_{0.94V_0}^{1.06V_0} [E^{\mathrm{a}}(V) - E^{\mathrm{b}}(V)]^2 dV}{0.12V_0}} \qquad (10)$$

где Ea и Eb обозначают энергии, вычисленные с использованием методов a и b соответственно.

На рисунке 5 (b) показаны значения ΔEOS различных моделей машинного обучения по отношению к эталонным данным DFT для различных элементарных систем, а также кривые EOS для этих ML-IAP. Во всех случаях ΔEOS для всех ML-IAP для всех элементов находится в пределах 2 мэВ / атом, что является порогом для «неразличимого EOS», ранее использовавшегося при оценке различных кодов DFT [87]. Примечательно, что, несмотря на относительно высокие ошибки прогноза моделей SNAP, представленных на рисунке 3 (a), они работают значительно лучше при прогнозировании кривых УС, со всеми значениями ΔEOS ниже 1 мэВ / атом для разных химических структур. Модели NNP незначительно отклоняются от кривых DFT как при растяжении, так и при деформации сжатия для систем с ГЦК, в то время как для алмазных систем отклонение моделей NNP от кривой DFT сопоставимо с таковыми для моделей GAP и MTP, что подтверждается сравнением калибровки Δ. В целом, дать высокоточные прогнозы EOS в алмазной системе сложнее, чем в системах с ГЦК и ОЦК. В дополнение к точности на уровне DFT в уравнениях предсказания состояния, предсказанные кривые фононной дисперсии для всех ML-IAP, исследованных в этой работе, превосходно согласуются с эталоном DFT (см. Рисунок S1 – рисунок S6 в дополнительной информации).





(б) Кривые зависимости энергии от объема

Кривые зависимости энергии от объема

РИС. 5. Оценка точности ML-IAP в прогнозировании уравнения состояния. (a) Сравнение калибровки Δ

обеспечивает количественную оценку отклонения между кривой EOS от каждого ML-IAP и кривой DFT.

(b) Кривые УС для всех шести элементов с использованием DFT и четырех ML-IAP.

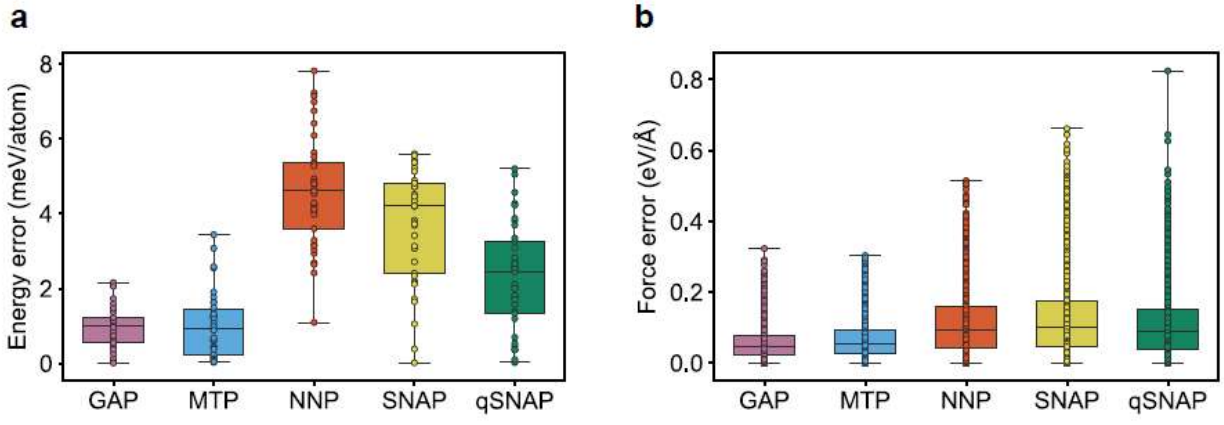


РИС. 6. Распределение ошибок в (а) предсказанных энергиях (б) предсказанных сил для выбранных структур из МД.

моделирования с использованием каждого ML-IAP. Прямоугольная рамка указывает межквартильный размах (IQR), в то время как линия в рамке указывает на медианное значение.

Д. Точность траекторий молекулярной динамики (МД)

Одно из основных приложений ML-IAP - это моделирование молекулярной динамики (МД). Чтобы оценить способность ML-IAP обеспечивать стабильные траектории MD, мы провели моделирование NV T MD при 1300 K (0,5 × температура плавления) на 3 × 3 × 3 54-атомной суперячейке объемного Mo в течение 0,25 нс с использованием LAMMPS. с разными ML-IAP. Затем было сделано 40 снимков с интервалом 2,5 пс для каждой MD-траектории, и для этих снимков были выполнены статические вычисления DFT. На рис.6 показано распределение ошибок по энергиям и силам отобранных структур. В соответствии с предыдущими результатами, модели GAP и MTP обычно показывают меньшие ошибки в энергиях и силах, чем модели NNP, SNAP и qSNAP. Модель GAP имеет не только самую низкую медиану, но и наименьший межквартильный диапазон (IQR) ошибок в энергиях и силах. Несколько интересно то, что модель NNP имеет более высокие ошибки энергии, но меньшие ошибки силы, чем SNAP и qSNAP. Для согласованности сравнения все модели, показанные здесь, являются «оптимальными» моделями, основанными на ~ 100 обучающих структурах. Вероятно, что более крупный обучающий набор улучшит производительность моделей NNP и qSNAP. (Рисунок 4).

F. Точность полиморфных разностей энергии

Чтобы оценить способность ML-IAP экстраполировать на невидимые данные, мы вычислили разницу в энергии между полиморфом основного состояния DFT и полиморфом с низкой энергией для каждого элемента, представленного на рисунке 7. Полиморфы с низкой энергией соответствуют к структурам ОЦК, ГЦК и вюрцита (гексагональный алмаз) для систем ГЦК, ОЦК и алмаза соответственно. Следует отметить, что при обучении ML-IAP использовались только структуры основного состояния, и эти низкоэнергетические полиморфы не присутствовали в обучающих структурах. За исключением Li, который имеет чрезвычайно небольшую разницу в энергии между ГЦК и ОЦК структурами в DFT, все ML-IAP способны качественно воспроизвести разницу в энергии между полиморфами. Для большинства систем ML-IAP способны воспроизводить энергетические различия между полиморфами с точностью до 10-20 мэВ / атом; Основным исключением является Mo, который показывает большую

разницу энергий между ГЦК- и ОЦК-структурами. Одно примечательное наблюдение заключается в том, что модель GAP показывает наибольшую ошибку в прогнозировании разницы в энергии между структурами вюрцита и алмаза в Si и Ge по сравнению с другими ML-IAP, несмотря на относительно низкую MAE в прогнозируемых энергиях в этих системах (см. Рис. (a)). Мы полагаем, что это может быть связано с тем, что GAP может быть более чувствительным к отсутствующим эталонным конфигурациям, в то время как другие IAP могут более эффективно экстраполировать взаимодействия на эту невидимую конфигурацию. Несколько удивительно, что линейная модель SNAP демонстрирует одну из лучших характеристик в воспроизведении полиморфных энергетических различий во всех системах, превосходя даже GAP и MTP для Mo, Si и Ge, несмотря на значительно более высокие MAE по энергиям и силам.

IV. ВЫВОДЫ

Мы выполнили всестороннюю объективную оценку моделей GAP, MTP, NNP, SNAP и qSNAP ML-IAP, используя последовательно сгенерированные данные DFT по шести элементным системам, охватывающим различные кристаллические структуры (ГЦК, ОЦК и алмаз), химический состав (металлы основной группы, переходные металлы и полупроводники) и связи (металлические и ковалентные). Эта оценка проводится по трем ключевым показателям, которые имеют решающее значение для любого потенциального пользователя этих ML-IAP:

1. Точность предсказанных энергий, сил и свойств как видимых, так и невидимых структур.
   туры;

   2. Требования к обучающим данным, которые влияют на количество дорогостоящих вычислений DFT, которые необходимо выполнить для обучения ML-IAP с заданной точностью; а также

   3. Вычислительные затраты, которые влияют на размер систем, в которых могут выполняться вычисления при заданном вычислительном бюджете.

   Эти три показателя неразрывно связаны - для всех четырех ML-IAP увеличение количества степеней свободы (с увеличением вычислительных затрат) и увеличение обучающих структур обычно приводит к более высокой точности. Мы демонстрируем применение границы Парето как средство определения оптимальных компромиссов между этими показателями. Мы обнаружили, что для всех ML-IAP существует «оптимальная» конфигурация, при которой дальнейшее расширение числа степеней свободы дает небольшое улучшение точности с увеличением вычислительных затрат.

Мы обнаружили, что все ML-IAP способны достигать точности, близкой к DFT, при прогнозировании энергии, сил и свойств материалов, существенно превосходя традиционные IAP. Модели GAP и MTP демонстрируют наименьшие МАЭ по энергиям и силам. Однако модели GAP являются одними из самых дорогих с точки зрения вычислений для данной точности (на основе текущих реализаций) и демонстрируют плохую экстраполяцию на полиморфы с более высокой энергией в алмазных системах. Действительно, простая линейная модель SNAP, которая имеет один из самых высоких значений MAE в предсказанных энергиях и силах, демонстрирует лучшую экстраполяцию к полиморфам с более высокими энергиями, а также воспроизводит уравнение состояния для алмазных систем. Модели NNP и qS-NAP показывают относительно высокие значения MAE в энергиях с небольшими размерами данных, но их можно смягчить с увеличением обучающих данных.

Еще один несколько неожиданный вывод заключается в том, что даже с относительно небольшими наборами обучающих данных из ~ 100-200 структур, модели GAP, MTP и SNAP, по-видимому, достаточно хорошо сходятся с точностью до мэВ атом-1 по энергии и 0,01 эВ A-1. точность в силах. В

Модели NNP и qSNAP можно улучшить с помощью больших наборов обучающих данных, но

МАЭ даже на ~ 100 структурах не являются чрезмерно высокими. Мы связываем эту производительность с процедурой генерации обучающих данных, которая направлена на выборку разнообразных структур как из основного состояния, так и из мультитемпературного моделирования AIMD. Другими словами, разнообразие обучающих данных, возможно, важнее количества.

Наконец, мы отмечаем, что одним из ограничений этого исследования является то, что мы не пытались объединить различные дескрипторы локальной среды (функции симметрии, SOAP, биспектр, тензоры моментов) с различными структурами машинного обучения (например, линейная регрессия в сравнении с гауссовским процессом). регрессия против нейронной сети). Выбор дескриптора влияет на то, насколько эффективно могут быть закодированы различные локальные среды, в то время как выбор структуры машинного обучения определяет функциональную гибкость в отображении отношений между дескрипторами и энергиями / силами. Выбор различных дескрипторов и моделей может дать лучший компромисс между точностью и стоимостью для конкретного приложения.

[1] W. Kohn and L. J. Sham, Physical Review **140**, A1133 (1965).

[2] L. J. Sham and M. Schlüter, Physical Review Letters **51**, 1888 (1983).

[3] I. Y. Zhang, X. Xu, Y. Jung, and W. A. Goddard, Proceedings of the National Academy of Sciences **108**, 19896 (2011).

[4] Y. Zhang, X. Xu, and W. A. Goddard, Proceedings of the National Academy of Sciences **106**, 4963 (2009).

[5] H. Ji, Y. Shao, W. A. Goddard, and Y. Jung, Journal of Chemical Theory and Computation **9**, 1971 (2013).

[6] A. K. Rappe, C. J. Casewit, K. S. Colwell, W. A. Goddard, and W. M. Skiff, Journal of the American Chemical Society **114**, 10024 (1992).

[7] S. L. Mayo, B. D. Olafson, and W. A. Goddard, The Journal of Physical Chemistry **94**, 8897 (1990).

[8] A. C. T. Van Duin, S. Dasgupta, F. Lorant, and W. A. Goddard, The Journal of Physical Chemistry A **105**, 9396 (2001).

[9] J. Behler, The Journal of Chemical Physics **145**, 170901 (2016), https://doi.org/10.1063/1.4966192.

[10] A. P. Bartók, R. Kondor, and G. Csányi, Physical Review B **87**, 184115 (2013).

[11] J. Behler and M. Parrinello, Physical Review Letters **98**, 146401 (2007).

[12] J. Behler, Physical Chemistry Chemical Physics **13**, 17930 (2011).

[13] W. J. Szlachta, A. P. Bartók, and G. Csányi, Physical Review B **90**, 104108 (2014).

[14] D. Dragoni, T. D. Daff, G. Csányi, and N. Marzari, Physical Review Materials **2**, 013808 (2018).

[15] A. Thompson, L. Swiler, C. Trott, S. Foiles, and G. Tucker, Journal of Computational Physics , 15 (2015).

[16] C. Chen, Z. Deng, R. Tran, H. Tang, I.-H. Chu, and S. P. Ong, Physical Review Materials **1**, 043603 (2017).

[17] X.-G. Li, C. Hu, C. Chen, Z. Deng, J. Luo, and S. P. Ong, Physical Review B **98**, 094104 (2018).

[18] Z. Deng, C. Chen, X.-G. Li, and S. P. Ong, npj Computational Materials **5**, 75 (2019).

[19] A. V. Shapeev, Multiscale Modeling & Simulation **14**, 1153 (2016).

[20] E. V. Podryabinkin and A. V. Shapeev, Computational Materials Science **140**, 171 (2017).

[21] K. Gubaev, E. V. Podryabinkin, G. L. W. Hart, and A. V. Shapeev, Computational Materials Science **156**, 148 (2019), 1806.10567.

[22] V. Botu and R. Ramprasad, International Journal of Quantum Chemistry **115**, 1074 (2015).

[23] V. Botu and R. Ramprasad, Physical Review B **92**, 094306 (2015).

[24] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, The Journal of Physical Chemistry C **121**, 511 (2017).

[25] I. Kruglov, O. Sergeev, A. Yanilkin, and A. R. Oganov, Scientific Reports **7**, 8512 (2017).

[26] Z. Li, J. R. Kermode, and A. De Vita, Physical Review Letters **114**, 096405 (2015).

[27] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, Science Advances **3**, e1603015 (2017).

[28] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, and K.-R. Müller, Nature Communications **8**, 872 (2017).

[29] M. Rupp, International Journal of Quantum Chemistry **115**, 1058 (2015).

[30] Y. Huang, J. Kang, W. A. Goddard, and L.-W. Wang, Physical Review B **99**, 064103 (2019).

[31] A. Shapeev, Computational Materials Science **139**, 26 (2017).

[32] H. Balamane, T. Halicioglu, and W. A. Tiller, Phys. Rev. B **46**, 2250 (1992).

[33] D. O'connor and J. Biersack, Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms **15**, 14 (1986).

[34] J. Godet, L. Pizzagalli, S. Brochard, and P. Beauchamp, Journal of Physics: Condensed Matter **15**, 6943 (2003).

[35] J. Behler, International Journal of Quantum Chemistry **115**, 1032 (2015).

[36] J. Behler, Angewandte Chemie International Edition **56**, 12828 (2017).

[37] F. C. Mocanu, K. Konstantinou, T. H. Lee, N. Bernstein, V. L. Deringer, G. Csányi, and S. R. Elliott, The Journal of Physical Chemistry B **122**, 8998 (2018).

[38] M. A. Wood, M. A. Cusentino, B. D. Wirth, and A. P. Thompson, arXiv:1902.09395 [cond-mat, physics:physics] (2019), arXiv:1902.09395 [cond-mat, physics:physics].

[39] J. Behler, The Journal of Chemical Physics **134**, 074106 (2011).

[40] N. Artrith and A. Urban, Computational Materials Science **114**, 135 (2016).

[41] T. Morawietz, A. Singraber, C. Dellago, and J. Behler, Proceedings of the National Academy of Sciences **113**, 8368 (2016).

[42] V. Quaranta, M. Hellstrom, and J. Behler, The journal of physical chemistry letters **8**, 1476 (2017).

[43] M. Eckhoff and J. Behler, Journal of chemical theory and computation (2019).

[44] N. Artrith, T. Morawietz, and J. Behler, Physical Review B **83**, 153101 (2011).

[45] W. Li, Y. Ando, E. Minamitani, and S. Watanabe, The Journal of Chemical Physics **147**, 214106 (2017).

[46] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Physical Review Letters **104**, 136403 (2010).

[47] V. L. Deringer, C. J. Pickard, and G. Csányi, Physical Review Letters **120**, 156001 (2018).

[48] V. L. Deringer and G. Csányi, Physical Review B **95**, 094203 (2017).

[49] P. Rowe, G. Csányi, D. Alfè, and A. Michaelides, Physical Review B **97**, 054303 (2018).

[50] A. P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, Physical Review X **8**, 041048 (2018).

[51] V. L. Deringer, N. Bernstein, A. P. Bartók, M. J. Cliffe, R. N. Kerber, L. E. Marbella, C. P. Grey, S. R. Elliott, and G. Csányi, The Journal of Physical Chemistry Letters **9**, 2879 (2018).

[52] M. A. Wood and A. P. Thompson, The Journal of Chemical Physics **148**, 241721 (2018).

[53] I. Novoselov, A. Yanilkin, A. Shapeev, and E. Podryabinkin, Computational Materials Science **164**, 46 (2019).

[54] E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, and A. R. Oganov, Physical Review B **99**, 064114 (2019), 1802.07605.

[55] I. S. Novikov, Y. V. Suleimanov, and A. V. Shapeev, Physical Chemistry Chemical Physics **20**, 29503 (2018), 1805.11924.

[56] M. de Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. Krishna Ande, S. van der Zwaag, J. J. Plata, C. Toher, S. Curtarolo, G. Ceder, K. A. Persson, and M. Asta, Scientific Data **2**, 150009 (2015).

[57] R. Tran, Z. Xu, B. Radhakrishnan, D. Winston, W. Sun, K. A. Persson, and S. P. Ong, Scientific Data **23**, 1 (2016).

[58] http://crystalium.materialsvirtuallab.org (2016), .

[59] G. Kresse and J. Furthmüller, Physical Review B **54**, 11169 (1996).

[60] P. E. Blöchl, Physical Review B **50**, 17953 (1994).

[61] J. P. Perdew, K. Burke, and M. Ernzerhof, Physical Review Letters **77**, 3865 (1996).

[62] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Computational Materials Science **68**, 314 (2013).

[63] A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G.-M. Rignanese, G. Hautier, D. Gunter, and K. A. Persson, Concurrency and Computation:

[64] B.-J. Lee, J.-H. Shim, and M. I. Baskes, Physical Review B **68**, 144112 (2003).

[65] S. M. Foiles, M. I. Baskes, and M. S. Daw, Physical Review B **33**, 7983 (1986).

[66] Y. Song, R. Yang, D. Li, W. T. Wu, and Z. X. Guo, Physical Review B **59**, 14220 (1999).

[67] J. B. Adams and S. M. Foiles, Physical Review B **41**, 3316 (1990).

[68] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, Physical Chemistry Chemical Physics **18**, 13754 (2016).

[69] M. I. Baskes, J. S. Nelson, and A. F. Wright, Physical Review B **40**, 6085 (1989).

[70] A. Singraber, J. Behler, and C. Dellago, Journal of Chemical Theory and Computation **15**, 1827 (2019).

[71] S. Plimpton, Journal of Computational Physics **117**, 1 (1995).

[72] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy (IEEE, Honolulu, HI, 2017) pp. 3296–3297.

[73] T. T. Nguyen, E. Székely, G. Imbalzano, J. Behler, G. Csányi, M. Ceriotti, A. W. Götz, and F. Paesani, The Journal of Chemical Physics **148**, 241725 (2018).

[74] B. Cheng, E. A. Engel, J. Behler, C. Dellago, and M. Ceriotti, Proceedings of the National Academy of Sciences **116**, 1110 (2019).

[75] X. W. Zhou, R. A. Johnson, and H. N. G. Wadley, Physical Review B **69**, 144113 (2004).

[76] A. Nichol and G. J. Ackland, Phys. Rev. B **93**, 184101 (2016).

[77] E. Asadi, M. Asle Zaeem, S. Nouranian, and M. I. Baskes, Acta Materialia **86**, 169 (2015).

[78] H. Park, M. R. Fellinger, T. J. Lenosky, W. W. Tipton, D. R. Trinkle, S. P. Rudin, C. Woodward, J. W. Wilkins, and R. G. Hennig, Physical Review B **85**, 214121 (2012).

[79] T. J. Lenosky, B. Sadigh, E. Alonso, V. V. Bulatov, T. D. de la Rubia, J. Kim, A. F. Voter, and J. D. Kress, Modelling and Simulation in Materials Science and Engineering **8**, 825 (2000).

[80] T. Kumagai, S. Izumi, S. Hara, and S. Sakai, Computational Materials Science **39**, 457 (2007).

[81] S. J. Mahdizadeh and G. Akhlamadi, Journal of Molecular Graphics and Modelling **72**, 1 (2017).

[82] G. Henkelman and H. Jónsson, The Journal of chemical physics **113**, 9978 (2000).

[83] R. Hill, Proceedings of the Physical Society. Section A **65**, 349 (1952).

[84] K. Lejaeghere, V. Van Speybroeck, G. Van Oost, and S. Cottenier, Critical Reviews in Solid State and Materials Sciences **39**, 1 (2014).

[85] G. I. Csonka, J. P. Perdew, A. Ruzsinszky, P. H. T. Philipsen, S. Lebègue, J. Paier, O. A. Vydrov, and J. G. Ángyán, Physical Review B **79**, 155107 (2009).

[86] V. N. Staroverov, G. E. Scuseria, J. Tao, and J. P. Perdew, Physical Review B **69**, 075102 (2004).

[87] K. Lejaeghere, G. Bihlmayer, T. Bjorkman, P. Blaha, S. Blugel, V. Blum, D. Caliste, I. E. Castelli, S. J. Clark, A. Dal Corso, S. de Gironcoli, T. Deutsch, J. K. Dewhurst, I. Di Marco, C. Draxl, M. Du ak, O. Eriksson, J. A. Flores-Livas, K. F. Garrity, L. Genovese, P. Giannozzi, M. Giantomassi, S. Goedecker, X. Gonze, O. Granas, E. K. U. Gross, A. Gulans, F. Gygi, D. R. Hamann, P. J. Hasnip, N. A. W. Holzwarth, D. Iu an, D. B. Jochym, F. Jollet, D. Jones, G. Kresse, K. Koepernik, E. Kucukbenli, Y. O. Kvashnin, I. L. M. Locht, S. Lubeck, M. Marsman, N. Marzari, U. Nitzsche, L. Nordstrom, T. Ozaki, L. Paulatto, C. J. Pickard, W. Poelmans, M. I. J. Probert, K. Refson, M. Richter, G.-M. Rignanese, S. Saha, M. Scheffler, M. Schlipf, K. Schwarz, S. Sharma, F. Tavazza, P. Thunstrom, A. Tkatchenko, M. Torrent, D. Vanderbilt, M. J. van Setten, V. Van Speybroeck, J. M. Wills, J. R. Yates, G.-X. Zhang, and S. Cottenier, Science **351**, aad3000 (2016).