

Abstract

Materials Informatics [1] refers to the idea that extensive amounts of data from the fields of Condensed Matter Physics and Materials Science, both experimental and computational, can be used to train machine-learning algorithms for materials design. In particular, examples are direct prediction of properties based on crystalline structures (Crystal Graph Convolutional Neural Networks) [2], improvement of the description accuracy and acceleration of molecular dynamics simulations (e.g., Neural Network Interatomic Potentials) [3], recognition of hidden data patterns in multi-dimensional datasets (Unsupervised Learning) [4], among other optimizations. The key success factor lies in the efficiency of the descriptors used to encode relevant information about properties of the system.

As a first point, in this Thesis, I present the idea of exploiting ab-initio electron charge density profiles from first-principles calculations in Machine Learning (ML) as effective descriptors, starting with the case of study of interstitial defects in bulk crystals [4]. Electron charge density-based descriptors arise from intrinsic interactions, therefore capturing the full landscape of electronic-related behaviors. I first apply this methodology to the context of hydrogen interstitial defects inside FCC crystals and the resulting induced perturbations in electron density fields. Radial Distribution Functions (RDFs) are built in the perturbed electron density fields and compared for a variety of defected FCC metal bulk crystals, exhibiting correlations of their characteristics with an extended dataset of properties, i.e. charge-density properties of hydrogen and metal atoms, structural and atomic properties of the undefected bulk crystals, including large-scale properties like melting point, heat of formation and boiling point. A major finding is the emergence of classes of charge responses, playing a crucial role in hydride composition search. Through unsupervised learning techniques, I demonstrate the possibility of building density-based composition maps classifying both the charge responses and hydrogen diffusive dynamics. Beyond classification, I also explore how to efficiently deploy electron charge density images in regression tasks, via Deep Learning approaches including multimodal transfer-learning on Electron Charge Density (ECDs) fields from bulk crystals [5]. I demonstrate the highly informative nature of representations extracted from visual (Convolutional Neural Networks on ECDs images) and semantic (Large Language Models on crystallographic textual information)

modalities, exploring a wide set of compositions for the regression of energetics- or structure-related properties. A comparison between a custom solution involving stacked modalities and ready-to-use multimodal models is presented. Such treatment shows potential impact on decision making for new low-cost AI methods in the field of Materials and Chemoinformatics.

As a second point, along the line of direct extraction of atomic representations with Neural Networks, I also investigate the predictive abilities of available foundational graph-based models in Materials Informatics. On the one hand of static structure-property predictions, I propose a substitutional alloying framework to test the transferrability of a pre-trained crystal graph convolutional neural-network model. The question of whether the learned representations of local atomic chemical environments *generalize* towards systems where model-known atomic species are contained in unknown-arrangements (out of training database) is a key question for Materials Discovery tasks, the ultimate scope for which such models have been developed [2]. On the other hand of dynamic simulations, graph-based models for interatomic potential development provide an opportunity to further explore the role of training configurations for accurate predictions. In particular, the importance of dataset building for such potentials is investigated for the modeling of hydrogen diffusion in magnesium through active-learning techniques, with a focus on fine-tuning of pre-trained universal deep-learning potentials [3]. The work reports remarkable agreement with the experimental results for hydrogen diffusion coefficients and transition barriers, at different temperatures and for different potentials.

Streszczenie

Informatyka materiałów [1] odnosi się do koncepcji wykorzystania obszernych ilości danych z dziedziny fizyki materii skondensowanej i nauki o materiałach, zarówno eksperymentalnych, jak w celu obliczeniowych, do trenowania algorytmów uczenia maszynowego w celu projektowania materiałów. Przykładowo, można wskazać bezpośrednie przewidywanie właściwości na podstawie struktur krystalicznych (Crystal Graph Convolutional Neural Networks) [2], poprawę dokładności opisu i przyspieszenie symulacji dynamiki molekularnej (np. Neural Network Interatomic Potentials) [3], rozpoznawanie ukrytych wzorców danych w wielowymiarowych zbiorach danych (Unsupervised Learning) [4], a także inne optymalizacje. Kluczowym czynnikiem sukcesu jest efektywność deskryptorów używanych do kodowania istotnych informacji o właściwościach systemu.

W pierwszym punkcie tej pracy przedstawiam ideę wykorzystania profili gęstości ładunku elektronowego obliczonych metodami pierwszych zasad w Machine Learning (ML) jako efektywnych deskryptorów, zaczynając od przypadku badania defektów międzywęzłowych w kryształach objętościowych [4]. Deskryptory oparte na gęstości ładunku elektronowego wynikają z wewnętrznych interakcji, co pozwala uchwycić pełen obraz zachowań związanych z elektronami. Na początek stosuję tę metodologię w kontekście defektów międzywęzłowych wodoru wewnątrz kryształów FCC i wynikających z nich perturbacji w polach gęstości elektronowej. Radial Distribution Functions (RDF) są tworzone w zaburzonych polach gęstości elektronowej i porównywane dla różnych kryształów metali FCC z defektami, wykazując korelacje ich cech z rozszerzonym zestawem właściwości, tj. właściwościami gęstości ładunku wodoru i atomów metalu, właściwościami strukturalnymi i atomowymi idealnych kryształów objętościowych, w tym właściwościami makroskopowymi, takimi jak temperatura topnienia, entalpia tworzenia i temperatura wrzenia. Kluczowym odkryciem jest wyłonienie się klas odpowiedzi ładunkowych, odgrywających kluczową rolę w poszukiwaniach składu wodorków. Za pomocą technik uczenia nienadzorowanego demonstruję możliwość budowy map składu opartych na gęstości, klasyfikujących zarówno odpowiedzi ładunkowe, jak i dynamikę dyfuzyjną wodoru. Oprócz klasyfikacji eksploruję także sposób efektywnego wykorzystania obrazów gęstości ładunku elektronowego w zadaniach regresji za pomocą podejść głębokiego uczenia, w tym multimodalnego transferu uczenia na

polach gęstości ładunku elektronowego (ECD, ang. electron charge density) w kryształach objętościowych [5]. Demonstruję bardzo informacyjny charakter reprezentacji wyekstrahowanych z modalności wizualnych (splotowe sieci neuronowe na obrazach ECD) i semantycznych (duże modele językowe na tekstowych informacjach krystalograficznych), eksplorując szeroki zestaw składów do regresji właściwości związanych z energetyką lub strukturą. Przedstawiono porównanie między niestandardowym rozwiązaniem złożonym z połączonych modalności a gotowymi modelami wielomodalnymi. Taka analiza ukazuje potencjalny wpływ na podejmowanie decyzji w zakresie nowych wydajnych metod AI w dziedzinie materiałoznawstwa i chemoinformatyki. W drugim punkcie, w ramach bezpośredniego pozyskiwania reprezentacji atomowych za pomocą sieci neuronowych, badam również zdolności predykcyjne dostępnych modeli podstawowych opartych na grafach w informatyce materiałowej. Z jednej strony, w kontekście statycznych przewidywań właściwości struktur, proponuję ramy zastępczego stopowania do testowania transferowości wstępnie wytrenowanego modelu krystalicznej splotowej sieci neuronowej. Kluczowym pytaniem dla zadania poszukiwania nowych materiałów (głównego celu, dla którego opracowano takie modele) jest to, czy wyuczone reprezentacje lokalnych chemicznych środowisk atomowych *generalizują się* na systemy, w których modelowane gatunki atomowe występują w nieznanach układach (poza bazą danych treningowych) [2]. Z drugiej strony, dynamiczne symulacje i modele oparte na grafach dla rozwoju potencjałów międzyatomowych umożliwiają dalsze badanie wpływu konfiguracji treningowych na dokładne przewidywania. W szczególności badam znaczenie budowy zestawu danych dla takich potencjałów w modelowaniu dyfuzji wodoru w magnezie za pomocą technik aktywnego uczenia, koncentrując się na dostrajaniu wstępnie wytrenowanych uniwersalnych potencjałów głębokiego uczenia [3]. Wyniki pracy wykazują znakomitą zgodność z wynikami eksperymentalnymi dotyczącymi współczynników dyfuzji wodoru i barier przejścia, w różnych temperaturach i dla różnych potencjałów.

Bibliography

- [1] Amir H. Naghdi, Dario Massa, Kamran Karimi, and Stefanos Papanikolaou. High entropy alloy composition design for mechanical properties. In Dr. Yu Yin, Prof. Han Huang, Dr. Mingxing Zhang, and Prof. Libo Zhou, editors, *High Entropy Alloys - Composition and Microstructure Design*, chapter 1. IntechOpen, Rijeka, 2024.
- [2] Dario Massa, Daniel Cieřliński, Amirhossein Naghdi, and Stefanos Papanikolaou. Substitutional alloying using crystal graph neural networks. *AIP Advances*, 14(1), 2024.
- [3] Andrea Angeletti, Luca Leoni, Dario Massa, Luca Pasquini, Stefanos Papanikolaou, and Cesare Franchini. Hydrogen diffusion in magnesium using machine learning potentials: a comparative study. (last review stages in npj comp. mat.). 2407.21088 arxiv, 2024.
- [4] Dario Massa, Efthimios Kaxiras, and Stefanos Papanikolaou. Alloy informatics through ab initio charge density profiles: Case study of hydrogen effects in face-centred cubic crystals. *Acta Materialia*, 268:119773, 2024.
- [5] Dario Massa, Grzegorz Kaszuba, Stefanos Papanikolaou, and Piotr Sankowski. Transfer learning in materials informatics: structure-property relationships through minimal but highly informative multimodal input. (accepted in neurips2024 - ml4ps workshop). 2401.09301 arxiv, 2024.