Studienprojekt: Modularisierung der Implementation des Klassifizierungsmodells *SynCoTrain*

Angelina Jellinek

Motivation

Die Erforschung von neuen Materialien ist eine fundamentale Zielstellung der modernen Wissenschaft. Diese ermöglichen beispielsweise Fortschritte im medizinischen Bereich und sind auch für Herausforderungen des Klimawandels relevant [1]. Die Beschleunigung der Entdeckung und Entwicklung neuer Materialien soll erreicht werden durch den Einsatz von Hochdurchsatzsimulationen und darauf folgenden gezielten Experimenten zum Suchen geeigneter Materialien mit wünschenswerten Eigenschaften [1, 2]. Damit ein Material mit gewünschten Eigenschaften überhaupt entwickelt und verwendet werden kann, muss es jedoch auch stabil und synthetisierbar sein [3].
Um die Stabilität vorherzusagen, wurden lange die Pauling-Regeln [4] oder die Kriterien für den Ladungsausgleich [5] eines Materials verwendet. Diese vereinfachten Ansätze haben sich jedoch als veraltet erwiesen [5, 6].
Materialwissenschaftler haben in neueren Versuchen daher oft die durch Dichtefunktionaltheorie (DFT) berechnete thermodynamische Stabilität als Stellvertreter für die Synthetisierbarkeit verwendet [3, 7]. Die Stabilität trägt zwar wesentlich zur Synthesefähigkeit bei, jedoch können Materialien auch unter alternativen thermodynamischen Bedingungen synthetisiert werden und anschließend durch kinetische Stabilisierung in der metastabilen Struktur bestehen bleiben [8]. Synthetisierbarkeit ist außerdem ein technologisches Problem. Materialien, die in der Theorie stabil wären, können erst synthetisiert werden, sobald eine praktische Methode für die Synthese gefunden wird. Da die Entscheidung, ob ein Material synthetisierbar ist, mit seiner Materialstruktur zusammenhängt und nicht durch eine einfache Formel zu bestimmen ist, kann hier maschinelles Lernen Anwendung finden [3].
Eine Herausforderung dabei ist die Verfügbarkeit von Materialstrukturdaten. In der Arbeit von Amariamir et al. [3] wurden die Experimentierdaten von der *Materials Project API*[1] abgefragt. Die unmarkierten Daten wurde von der *Open Quantum Materials Database*[2] heruntergeladen. Anschließend werden Graph Convolutional Neural Networks (GCNNs) verwendet, um mit den Daten zu lernen. Eine andere Möglichkeit wäre die [9, 10] Verwendung von Bildern von Kristallstrukturen oder eine Netzwerkanalyse der Zeitachse der Materialentdeckung im Hinblick auf ihre Stabilität [11]. GCNNs verarbeiten verglichen mit den zuvor genannten Methoden mehr Informationen, sind allerdings auch komplizierter in der Implementierung [3, 12].
Die zweite Herausforderung besteht in der fehlenden Verfügbarkeit von negativen Daten, die man typischerweise für eine Klassifizierungsaufgabe verwenden würde. Das liegt daran, dass fehlgeschlagene Syntheseversuche aus verschiedenen Gründen oft nicht öffentlich gemacht werden, da diese von vielen Faktoren wie beispielsweise den technologischen Möglichkeiten abhängen können [3].
Die dritte Herausforderung ergibt sich aus dem grundlegenden Aspekt, dass Modelle maschinellen Lernens immer einen gewissen Grad an Bias aufweisen, was auch bei der Vorhersage

[1] www.materialsproject.org
[2] https://oqmd.org/

der Synthetisierbarkeit von Materialien deutlich wird [3].
Die fehlenden negativen Daten in Kombination mit dem Bias machen eine genaue Abschätzung der Synthetisierbarkeit schwierig. Diese lässt sich beispielsweise durch die Nutzung mehrerer Modelle verbessern [3, 13].
Amariamir et al. [3] haben 2024 das halbüberwachte (semi-supervised) Klassifizierungsmodell *SynCoTrain* entworfen, um die Herausforderungen der Erforschung neuer Materialien und ihrer Synthetisierbarkeit zu bewältigen. Sie setzen Co-Training und PU-Learning ein. Co-Training ist ein iterativer, halbüberwachter Lernprozess, der für Szenarien mit einigen positiven Daten und einer Menge unmarkierter Daten entwickelt wurde [14, 15]. Bei jedem Schritt des Co-Trainings lernt *SynCoTrain* durch die von Mordelet und Vert eingeführte Methode des positiven und unmarkierten maschinellen Lernens (PU-Learning) [16]. Es wird mit zwei verschiedenen GCNN-Klassifikatoren trainiert, um trotz des Fehlens negativer Daten in der Trainingsmenge Abschätzungen der Synthetisierbarkeit zu treffen. Amariamir et al. [3] zeigen hiermit das Potenzial der Kombination von Co-Training und PU-Learning.

Ziel

In meinem Studienprojekt möchte ich die starre Implementation[3] des Verfahrens von Amariamir et al. [3] neu implementieren und für weitere Anwendungen flexibler einsetzbar machen. Hierfür soll der Python-Code modularisiert werden, sodass ohne viel Aufwand andere Modelle eingesetzt werden können.

Vorgehen

In den folgenden Abschnitten, werde ich die Arbeitsschritte des Studienprojekts erläutern. Hierbei gehe ich darauf ein, was in jedem Arbeitsschritt getan werden muss.

Bisherige Implementation nachvollziehen.
Der bisherige Stand der Implementierung von Amariamir et al. [3] muss auf dem Uniserver zum Laufen gebracht werden. Ziel ist es, den Programmcode nachzuvollziehen, um anschließend an ihm arbeiten zu können. Hierfür werden Teile des Algorithmus mit den von Amariamir et al. genutzten Eingabedaten erneut ausgeführt. Die vollständige Reproduktion der Experimente würde allerdings mehrere Wochen dauern.

Umstrukturierung des Programmcodes planen.
Um einen guten Überblick zu bekommen, soll einen Konzept entworfen werden, wie eine sinnvolle Modularisierung erreicht werden kann. Hierfür können beispielsweise Klassendiagramme nützlich werden, um die Entwicklung eines Pythonpakets zu planen. Ein erster Entwurf eines Zielklassendiagramms ist in Abbildung 1 zu sehen. Die neue Implementation soll mit Hilfe des *Strategy Pattern* [17] umgesetzt werden.

Überarbeitung des Programmcodes.
Schließlich wird der Programmcode überarbeitet. Hierbei liegt der Fokus auf Modularisierung. Zusätzlich wird geprüft, ob Teile der Implementierung im Sinne von Lesbarkeit umgeschrieben werden können.

[3] https://github.com/sasanamari/SynCoTrain

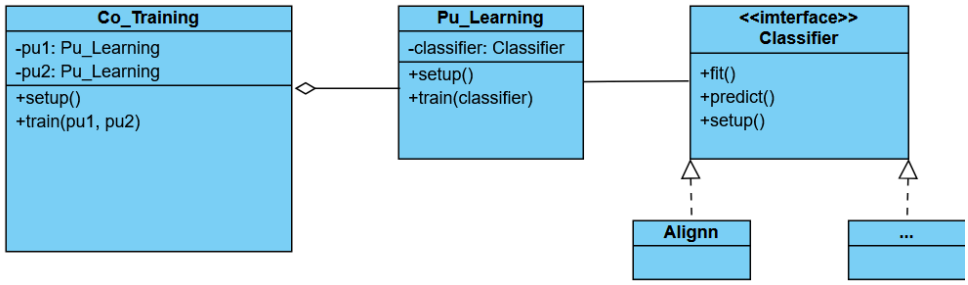

Abbildung 1: Konzeptionelles Klassendiagramm

Testskript erstellen.
Anschließend soll die Funktionalität ~~mit einfachen Klassifizierern~~ getestet werden. Auf einem kleinen Datensatz soll untersucht werden, ob SynCoTrain besser ist als PU-Learning alleine.

Related Work

Linus Pauling [4] veröffentlichte 1929 fünf Regeln, um Aussagen über die Struktur von ionisch aufgebauten Kristallen zu treffen. Diese bildeten lange eine wichtige Grundlage der Erforschung der chemischen Zusammensetzung und physikalischen Eigenschaften von kristallinen Stoffen [18]. 2020 zeigten George et al. [6] jedoch, dass die Regeln für die meisten Oxide schlecht funktionieren und als Kriterien für die Synthetisierbarkeit nicht geeignet sind.
Ein anderer Ansatz ist die Verwendung der durch Dichtefunktionaltheorie (DFT) berechneten thermodynamischen Stabilität [3, 7]. Sun et al. [8] führten hierzu 2016 eine groß angelegte Data-Mining-Studie auf einer Hochdurchsatz-Datenbank durch und konnten so neue physikalische Erkenntnisse gewinnen, die bei der Entwicklung metastabiler Materialien hilfreich sein können.
Da bei der Betrachtung der thermodynamischen Stabilität jedoch die kinetische Stabilisierung metastabiler Strukturen außer Acht gelassen wird, verwendet man für die genauere Abschätzung der Synthetisierbarkeit aufgrund der Komplexität dieser Aufgabe heute zunehmend maschinelles Lernen [3].
Antoniuk et al. [5] untersuchten 2023 die Nutzung des Deep-Learning-Klassifizierungsmodells *SynthNN* und konnten so die DFT-basierten Berechnungen in Präzision weit übertreffen.
Mit Graph Convolutional Neural Networks (GCNNs) lassen sich besonders viele Informationen über die Materialstruktur kodieren und daraus lernen [3]. GCNNs wurden unter anderem 2020 von Jang et al. [19] verwendet. Das von Ihnen entwickelte maschinelle Lernmodell zur Quantifizierung der Synthesewahrscheinlichkeit basiert auf PU-Learning.
2010 entwickelten Mordelet und Vert [16] eine konzeptionell einfache Methode des PU-Learnings mit welcher sie gute und schnelle Ergebnisse erzielen konnten.
2024 kombinierten Amariamir et al. [3] den Ansatz von Mordelet und Vert [16] mit Co-Training, welches für Szenarien mit einer großen Menge unmarkierten Daten entwickelt wurde. In ihrer Arbeit „SynCoTrain: Predicting synthesizability using Co-Training and PU-Learning“ stellen sie das halbüberwachte Klassifizierungsmodell *SynCoTrain* vor. Sie zeigten, dass die Verwendung von GCNNs und mehrerer PU-Learner als Bausteine für das Co-Training die Zuverlässigkeit und Genauigkeit der Vorhersagen verbessern kann. Die Implementation *SynCoTrains* [4] von

[4] https://github.com/sasanamari/SynCoTrain

Amariamir et al. [3] bildet die Grundlage für dieses Studienprojekt.

Literatur

[1] Juan de Pablo, Nicholas Jackson, Michael Webb, Long-Qing Chen, Joel Moore, Dane Morgan, Ryan Jacobs, Tresa Pollock, Darrell Schlom, Eric Toberer, James Analytis, Ismaila Dabo, Dean DeLongchamp, Gregory Fiete, Gregory Grason, Geoffroy Hautier, Yifei Mo, Krishna Rajan, Evan Reed, and Ji-Cheng Zhao. New frontiers for the materials genome initiative. *npj Computational Materials*, 5:41, 04 2019.

[2] John Rodgers. Materials informatics. *MRS Bulletin*, 31:975 – 980, 12 2006.

[3] S. Amariamir, J. George, and P. Benner. SynCoTrain: Predicting synthesizability using Co-Training and PU-Learning, to be published.

[4] Linus Pauling. The principles determining the structure of complex ionic crystals. *Journal of the American Chemical Society*, 51(4):1010–1026, 1929.

[5] Evan R Antoniuk, Gowoon Cheon, George Wang, Daniel Bernstein, William Cai, and Evan J Reed. Predicting the synthesizability of crystalline inorganic materials from the data of known material compositions. *npj Computational Materials*, 9(1):155, 2023.

[6] Janine George, David Waroquiers, Davide Di Stefano, Guido Petretto, G.-M Rignanese, and Geoffroy Hautier. The limited predictive power of the pauling rules. *Angewandte Chemie International Edition*, 59, 03 2020.

[7] Walter Kohn and L Sham. Density functional theory. In *Conference Proceedings-Italian Physical Society*, volume 49, pages 561–572. Editrice Compositori, 1996.

[8] Wenhao Sun, Stephen Dacek, Shyue Ong, Geoffroy Hautier, Anubhav Jain, William Richards, Anthony Gamst, Kristin Persson, and Gerbrand Ceder. The thermodynamic scale of inorganic crystalline metastability. *Science Advances*, 2:e1600225–e1600225, 11 2016.

[9] Andrew Lee, Suchismita Sarker, James E Saal, Logan Ward, Christopher Borg, Apurva Mehta, and Christopher Wolverton. Machine learned synthesizability predictions aided by density functional theory. *Communications Materials*, 3(1):73, 2022.

[10] Fleur Legrain, Jesús Carrete, Ambroise van Roekeghem, Georg K.H. Madsen, and Natalio Mingo. Materials screening for the discovery of new half-heuslers: Machine learning versus ab initio methods. *The Journal of Physical Chemistry B*, 122(2):625–632, 2018. PMID: 28742351.

[11] Muratahan Aykol, Vinay I. Hegde, Linda Hung, Santosh K. Suram, Patrick K. Herring, Christopher M. Wolverton, and Jens S. Hummelshøj. Network analysis of synthesizable materials discovery. *Nature Communications*, 10, 2018.

[12] Kanishka Singh, Jannes Münchmeyer, Leon Weber, Ulf Leser, and Annika Bande. Graph neural networks for learning molecular excitation spectra. *Journal of Chemical Theory and Computation*, 18(7):4408–4417, 2022. PMID: 35671364.

[13] Kangming Li, Brian DeCost, Kamal Choudhary, Michael Greenwood, and Jason Hattrick-Simpers. A critical examination of robustness and generalizability of machine learning prediction of materials properties. *npj Computational Materials*, 9(1):55, 2023.

[14] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, page 92–100, New York, NY, USA, 1998. Association for Computing Machinery.

[15] François Denis, Anne Laurent, and Marc Tommasi. Text classification and co-training from positive and unlabeled examples. *Proceedings of the ICML 2003 Workshop: The Continuum from Labeled to Unlabeled Data*, 01 2003.

[16] Fantine Mordelet and Jean-Philippe Vert. A bagging svm to learn from positive and unlabeled examples. *Pattern Recognit. Lett.*, 37:201–209, 2010.

[17] Erich Gamma. Design patterns: elements of reusable object-oriented software. *Person*

Education Inc, 1995.

[18] H. Salmang, R. Telle, and H. Scholze. *Keramik*. Number Bd. 1. Springer Berlin Heidelberg, 2006.

[19] Jidon Jang, Geun Ho Gu, Juhwan Noh, Juhwan Kim, and Yousung Jung. Structure-based synthesizability prediction of crystals using partially supervised learning. *Journal of the American Chemical Society*, 142(44):18836–18843, 2020. PMID: 33104335.