УДК 004.822+89

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ С ИСПОЛЬЗОВАНИЕМ ДАННЫХ OPENMX И AFLOW ДЛЯ ОПТИМИЗАЦИИ ВЫЧИСЛИТЕЛЬНОГО ЭКСПЕРИМЕНТА

Н.А. Шиманский
Соискатель кафедры физики твердого тела и нанотехнологий физического факультета БГУ, системный архитектор компании Andersen Lab (ПВТ)
nikita.shymanski@gmail.com

Н.А. Шиманский

Окончил Белорусский государственный университет. Занимается проектированием и разработкой IT-решений в области бизнеса и науки. Область научных интересов – разработка программных средств для оптимизации решения прикладных задач в области наноматериаловедения с применением Machine Learning & Generative AI.

Аннотация: В работе предлагается метод ускорения вычислительного моделирования наноматериалов за счёт предварительного отбора структур с помощью нейросетевых моделей. Обучающая выборка формируется на базе открытых данных кристаллических структур из базы AFLOW и включает следующие параметры: геометрические характеристики решётки, координаты атомов и рассчитанные значения энергии образования. Для построения классификатора, прогнозирующего тип решётки Браве и энергию образования новых виртуальных конфигураций (генерируемых на основе входных файлов программного пакета OpenMX), применены методы градиентного бустинга (XGBoost, LightGBM) и архитектуры глубоких нейронных сетей. Предложенный подход позволяет исключить заведомо нестабильные структуры из последующего цикла ресурсоёмких расчётов методом теории функционала плотности. Результаты моделирования интегрируются с ранее созданной платформой Agentic AI, что обеспечивает автоматизированное пополнение базы знаний и валидацию полученных предсказаний.

Ключевые слова: машинное обучение, нейронные сети, XGBoost, OpenMX, AFLOW, наноматериалы, 2D-структуры.

Введение. Современное материаловедение характеризуется возрастающей ролью вычислительных методов, обеспечивающих прогнозирование физико-химических свойств новых соединений без проведения дорогостоящих экспериментальных исследований. Программные комплексы, реализующие теорию функционала плотности (DFT), например OpenMX [1–3], позволяют с высокой точностью рассчитывать электронную структуру материалов. Тем не менее, одной из ключевых проблем остаётся трудоёмкость подготовки входных данных и перебор значительного числа возможных структурных конфигураций, включающих вариации параметров кристаллической решётки, атомных позиций и наличия разных типов дефектов. Даже при задействовании высокопроизводительных вычислительных кластеров полный перебор всех потенциальных вариантов конфигураций сопряжён с существенными временными затратами, достигающими нескольких лет машинного времени. Ранее авторами были предложены методы автоматизации обработки экспериментальных данных с применением больших языковых моделей (LLM) и агентных систем [4–7], позволяющие эффективно собирать и структурировать информацию из открытых источников. Однако задача быстрого отбора перспективных структур для DFT-моделирования оставалась открытой. В данной работе представлена предиктивная модель на основе методов машинного обучения, обученная на данных базы AFLOW [8]. Модель обеспечивает отсеивание нефизичных конфигураций уже на этапе генерации виртуальных кристаллов, снижая тем самым вычислительную нагрузку на последующие DFT-расчёты.

Постановка задачи. Эффективность компьютерного моделирования в значительной степени определяется корректностью исходных предположений о структурной организации материала. Для наноматериалов на основе оксидов и халькогенидов молибдена (включая MoS_2, MoO_2, MoO_3 и родственные соединения) вариация параметров кристаллической решётки, взаимного расположения слоёв и атомных позиций формирует многомерное пространство поиска, содержащее миллионы потенциальных конфигураций. Значительная часть этих конфигураций не удовлетворяет критериям физической реализуемости: они либо не соответствуют ни одной из 14 решёток Браве, либо характеризуются положительной энергией образования, что свидетельствует о термодинамической нестабильности системы. Выполнение прямых расчётов для каждой такой конфигурации с использованием программного пакета OpenMX является вычислительно нецелесообразным. В связи с этим актуальной задачей выступает разработка предиктивного инструмента, способного на основе «сырых» структурных параметров (аналогичных тем, что используются во входных файлах OpenMX) оперативно классифицировать структуры по типу решётки Браве и оценивать их энергию образования. Дополнительно требуется обеспечить интеграцию данного инструмента с ранее разработанной платформой Agentic AI на базе AWS Bedrock [7], что позволит автоматизировать пополнение обучающей выборки и верификацию расчётных результатов.

Методология и архитектура. Основой предлагаемого методологического подхода служит формирование репрезентативной обучающей выборки на базе данных открытой базы AFLOW [8], содержащей результаты DFT-релаксации для тысяч кристаллических соединений. Для каждого материала из базы извлекаются следующие структурные и энергетические характеристики: геометрические параметры элементарной ячейки (a, b, c, α, β, γ); координаты атомов в элементарной ячейке; тип решётки Браве (в качестве категориальной целевой метки); энергия образования. Для преобразования атомных координат в набор признаков, инвариантных относительно пространственных трансляций и вращений, применяются функции радиального распределения (RDF) и угловые функции распределения (ADF). Указанные функции вычисляются для всех значимых пар химических элементов в структуре (Mo–Mo, Mo–S, S–S и т. д.). Полученный вектор признаков дополняется рядом глобальных структурных характеристик, включая объём

элементарной ячейки и отношение длин её сторон, что позволяет учесть макроскопические особенности кристаллической структуры.

В рамках исследования определены две целевые переменные, соответствующие различным типам задач машинного обучения: тип решётки Браве (14 классов) – задача многоклассовой классификации; энергия образования (непрерывная величина) – задача регрессии. Для совместного прогнозирования указанных целевых переменных рассматриваются два альтернативных подхода. Первый включает использование двух отдельных моделей градиентного бустинга: модель XGBoost [9] применяется для решения задачи классификации (определение типа решётки Браве), а модель LightGBM – для регрессионного прогнозирования энергии образования. Второй подразумевает применение единой многозадачной нейронной сети (Multi-Task Learning) на базе фреймворков Keras/TensorFlow. Архитектура сети предусматривает общий скрытый блок, состоящий из трёх полносвязных слоёв с регуляризацией методом Dropout, и два специализированных выходных слоя: слой с функцией активации Softmax – для многоклассовой классификации типов решётки Браве; линейный выходной слой – для регрессионной оценки энергии образования. Предложенный подход позволяет одновременно учитывать общие закономерности в данных, релевантные для обеих задач, и специфические особенности каждой целевой переменной.

Обучение модели выполняется на выборке, сбалансированной по классам решёток Браве, при недостаточной представленности отдельных классов применяется метод преддискретизации (oversampling) для устранения дисбаланса. Для комплексной оценки качества модели используются следующие метрики: точность (accuracy) и F1-мера – для оценки эффективности решения задачи классификации (определение типа решётки Браве), и средняя абсолютная ошибка (MAE) – для количественной оценки точности регрессионного прогноза энергии образования.

На этапе практического применения обученная модель обрабатывает сгенерированные виртуальные структуры, создаваемые путём малых смещений атомов в известных кристаллических фазах. Структуры, для которых модель предсказывает низкосимметричный тип решётки (триклинную или моноклинную систему) и/или положительную энергию образования свыше 0,1 эВ/атом автоматически исключаются как неперспективные. Оставшиеся конфигурации, удовлетворяющие критериям структурной симметрии и термодинамической стабильности, направляются на полноценный расчёт методом теории функционала плотности (DFT) с использованием программного пакета OpenMX.

Ключевым элементом разработанного подхода выступает интеграция с ранее созданной платформой Agentic AI [7], базирующейся на фундаментальной языковой модели Anthropic Claude 4.6 Sonnet/Opus и механизме расширенного извлечения знаний (Retrieval-Augmented Generation, RAG). Данная платформа обеспечивает автоматизированное пополнение базы знаний актуальными научными публикациями, а также структурированное извлечение из них данных о структурных параметрах и физико-химических свойствах материалов с последующим пополнением обучающей выборки новыми репрезентативными примерами. Важной функцией платформы является реализация процедур валидации предсказаний. Для ограниченного подмножества структур, классифицированных моделью как «физичные», то есть соответствующих критериям физической реализуемости, инициируются контрольные расчёты методом теории функционала плотности (DFT) в программном пакете OpenMX. Результаты этих расчётов сопоставляются с прогнозными значениями модели, что позволяет не только верифицировать точность предсказаний и выявлять случаи расхождения между прогнозом и расчётными данными, но и непрерывно корректировать, а также улучшать качество фильтрации виртуальных структур на последующих итерациях обучения. Таким образом

формируется замкнутый цикл самообучения, обеспечивающий постепенное повышение точности предиктивной модели.

На рисунке 1 представлена обобщённая архитектура решения, включающая модули сбора данных из AFLOW, генерации виртуальных структур, обучения модели фильтрации и интеграции с Agentic AI.

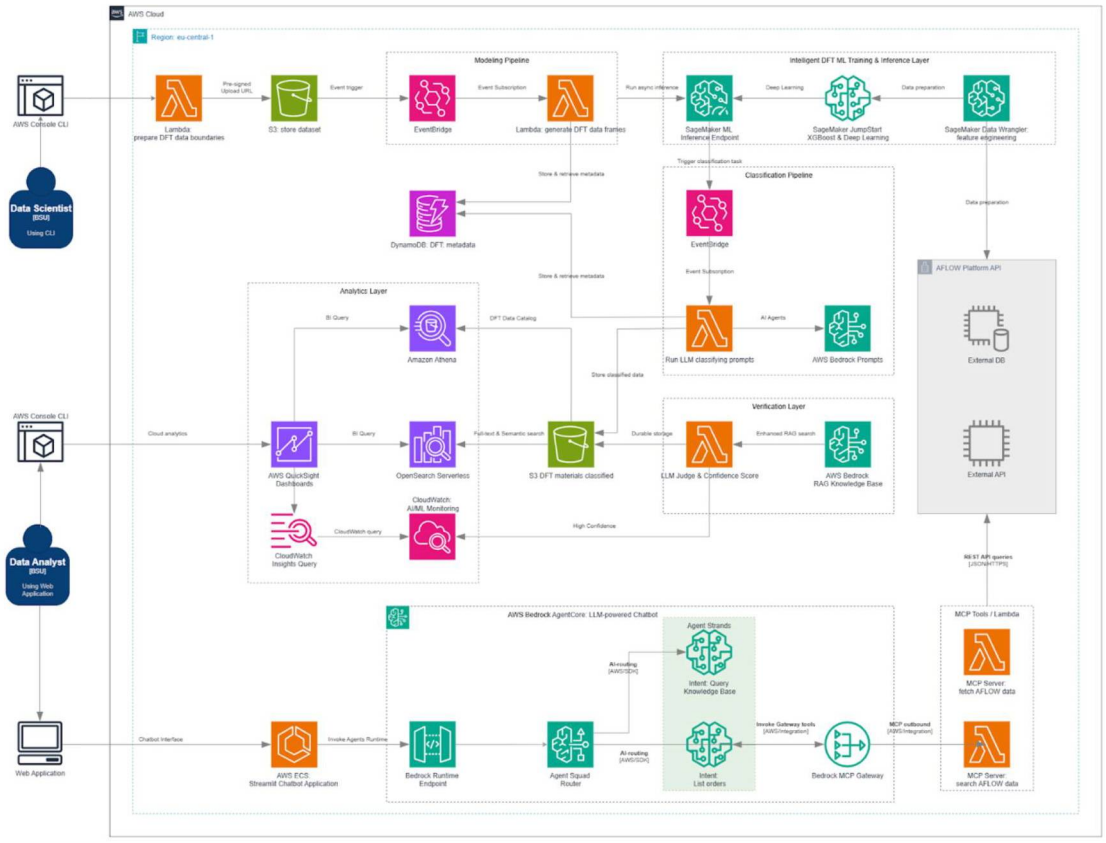

Рисунок 1. Целевая архитектура облачной системы предиктивной фильтрации структур наноматериалов с использованием машинного обучения и платформы Agentic AI

Программная реализация разработанного решения выполнена на языке программирования Python применением библиотек ИИ-агентов Agent Strands SDK и библиотек машинного обучения XGBoost, LightGBM и TensorFlow/Keras. Для доступа к данным открытой базы AFLOW использован интерфейс REST API, обеспечивающий структурированный обмен информацией и извлечение необходимых структурных и энергетических параметров кристаллических соединений. Обучение и развёртывание моделей машинного обучения осуществлено с использованием облачного сервиса AWS SageMaker. Это позволило автоматизировать ключевые этапы вычислительного конвейера: подготовку данных посредством инструмента SageMaker Data Wrangler и выполнение инференса через компонент SageMaker ML Inference. Агентная часть системы реализована на базе платформы AWS Bedrock AgentCore Runtime в сочетании с бессерверными функциями AWS Lambda, что обеспечивает гибкое управление вычислительными процессами и интеграцию разнородных компонентов системы. Данная архитектура подробно описана в предыдущей работе [7].

Апробация подхода выполнена на соединениях семейства Mo–S–O. Сгенерировано 105 структурных вариантов на основе MoS_2 и MoO_3 с вариацией параметров решётки в пределах ±5% и случайными смещениями атомов до 0,1 Å. Модель XGBoost (300-400 деревьев) отсеяла 94 % конфигураций как нестабильные. Для оставшихся 600 структур

выполнены контрольные расчёты в OpenMX. Ошибка предсказания энергии образования составила 0,08 эВ/атом, точность определения решётки Браве – 91 %. В 97 % случаев структуры, классифицированные как «физичные», демонстрировали сходимость SCF-цикла и отрицательную энергию образования.

Заключение. Предложенный метод предварительного отбора структур на основе машинного обучения позволяет на порядки сократить временные затраты на DFT-моделирование наноматериалов. Использование данных базы AFLOW обеспечивает формирование репрезентативной обучающей выборки, а комбинированное применение методов градиентного бустинга и глубоких нейронных сетей гарантирует высокую точность прогнозирования ключевых критериев физической реализуемости структур – типа решётки Браве и энергии образования. Интеграция с платформой Agentic AI создаёт возможность непрерывного пополнения базы знаний и автоматической валидации результатов, что существенно повышает адаптивность системы к новым классам материалов. Разработанный подход потенциально применим к широкому спектру наноструктур, включая гетеропереходы и дефектные системы, что будет являться предметом дальнейших исследований.

Список литературы

[1] Ozaki, T. Variationally optimized atomic orbitals for large-scale electronic structures / T. Ozaki // Phys. Rev. B. 2003. Vol. 67. P. 155108.

[2] Ozaki, T. Numerical atomic basis orbitals from H to Kr / T. Ozaki, H. Kino // Phys. Rev. B: Condens. Matter Mater. Phys. 2004. Vol. 69. P. 195113.

[3] Ozaki, T. Efficient projector expansion for the ab initio LCAO method / T. Ozaki, H. Kino // Phys. Rev. B. 2005. Vol. 72. P. 045121.

[4] Шиманский, Н.А. Автоматизация обработки результатов исследования структуры и свойств наноматериалов / Н.А. Шиманский, А.В. Баглов, Л.С. Хорошко // BIG DATA и анализ высокого уровня = BIG DATA and Advanced Analytics : сборник научных статей IX Международной научно-практической конференции, Минск, 17–18 мая 2023 г. : в 2 ч. Ч. 1 / Белорусский государственный университет информатики и радиоэлектроники ; редкол.: В. А. Богуш [и др.]. – Минск, 2023. – С. 296–300.

[5] Шиманский, Н. А. Автоматизация обработки результатов исследования структуры материалов / Н.А. Шиманский, А.В. Баглов, Л.С. Хорошко // Information Tehnologies and Systems 2023 (ITS 2023) : материалы международной научной конференции, Минск, Беларусь, 22 ноября / ред. Л. Ю. Шилин [и др.]. – Минск : БГУИР, 2023. – С. 207.

[6] Шиманский, Н.А. Автоматизация обработки результатов исследования структуры наноматериалов с использованием методов BIG DATA & MACHINE LEARNING / Н.А. Шиманский, А.В. Баглов // Математические методы и компьютерное моделирование в ФКС. – Гродно: ГрГУ, 2024. – С. 159.

[7] Шиманский, Н.А. Экспресс-анализ структурных и электронных свойств наноматериалов методами Big Data, Large Language Models & Generative AI / Н.А. Шиманский, А.В. Баглов, Л.С. Хорошко // Компьютерное проектирование в электронике = Electronic Design Automation : сб. трудов Междунар. науч.-практ. конф. (Республика Беларусь, г. Минск, 28 ноября 2024 г.) / редкол. : В. Р. Стемпицкий [и др.]. – Минск : БГУИР, 2024. – С. 100–103.

[8] Curtarolo, S. et al. AFLOW: An automatic framework for high-throughput materials discovery / S. Curtarolo et al. // Comput. Mater. Sci. 2012. Vol. 58. P. 218–226.

[9] Nemeth, M. The Comparison of Machine-Learning Methods XGBoost and LightGBM to Predict Energy Development / M. Nemeth, D. Borkin, G. Michalconok // In: Computational Statistics and Mathematical Modeling Methods in Intelligent Systems. CoMeSySo 2019. Advances in Intelligent Systems and Computing / Springer Cham. 2019. Vol 1047. P.208–215.

Авторский вклад

Шиманский Никита Андреевич – концептуализация метода предиктивной фильтрации, разработка моделей машинного обучения, реализация интеграции с платформой Agentic AI, написание программного кода, подготовка текста статьи.

Автор выражает благодарность своему научному руководителю канд. физ.-мат. наук, доценту Хорошко Л.С. за помощь в постановке задачи, выбор и обоснование методов и общее руководство работой, а также Баглову А.В. за техническую помощь в сборе и анализе данных для данной публикации.

PREDICTIVE FILTERING OF NANOMATERIALS STRUCTURES USING MACHINE LEARNING METHODS BASED ON OPENMX AND AFLOW DATA

N.A. Shymanski
Postgraduate of Department of Solid State Physics and Nanotechnologies, Faculty of Physics, BSU;
Solutions Architect of Andersen Lab (HTP)

Abstract. The paper proposes a method to accelerate computational modeling of nanomaterials by preliminary filtering of structures using neural network models. Based on open crystal structure databases (AFLOW), a training set containing lattice parameters, atomic positions, and calculated formation energies is formed. Using gradient boosting methods (XGBoost, LightGBM) and deep neural networks, a classifier is built that predicts the Bravais lattice type and formation energy for new virtual configurations generated from OpenMX input files. This allows rejecting obviously unstable structures before resource-intensive DFT calculations. The obtained results are integrated with the previously developed Agentic AI platform for automated knowledge base enrichment and prediction validation.

Keywords: machine learning, neural networks, XGBoost, OpenMX, AFLOW, predictive filtering, nanomaterials, 2D structures.