УДК 004.822+89

СТАТИСТИЧЕСКИЙ АНАЛИЗ И МОДЕЛИРОВАНИЕ СВОЙСТВ НАНОМАТЕРИАЛОВ МЕТОДАМИ LARGE LANGUAGE MODELS, AGENTIC AI & MACHINE LEARNING

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

А.В. Баглов
Старший научный сотрудник НИЛ энергоэффективных материалов и технологий физического факультета БГУ
baglov@bsu.by

Л.С. Хорошко
Ведущий научный сотрудник НИЛ энергоэффективных материалов и технологий физического факультета БГУ, канд. физ.-мат. наук, доцент
khoroshko@bsu.by

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.

А.В. Баглов
Окончил Белорусский государственный университет информатики и радиоэлектроники. Проводит научные исследования в рамках компьютерного моделирования электронной структуры перспективных материалов.

Л.С. Хорошко
Окончила Белорусский государственный университет информатики и радиоэлектроники. Область научных интересов связана с разработкой наноразмерных катализаторов фотостимулированных реакций и управляемым синтезом наноматериалов.

Аннотация: В данной работе рассматривается возможность комбинированного использования генеративного машинного обучения и классических нейронных сетей в качестве методов работы с большими данными для предиктивного анализа электронных свойств полупроводниковых наноматериалов, не ограничивая при этом общность данного подхода для иных кристаллических материалов и гетероструктур.

Ключевые слова: машинное обучение, нейронные сети, большие языковые модели, большие данные, наноматериалы, статистический анализ, предсказательный анализ.

Введение. Успешное развитие современного материаловедения, в особенности, в области наноматериалов, практически невозможно без использования вычислительных средств и методов компьютерной автоматизации. Стремительное развитие нейронных сетей и программных интерфейсов для них (*Discriminative AI, Generative AI*) позволяет анализировать в кратчайшие сроки значительные объемы экспериментальных данных и не только исследовать, но и с высокой достоверностью предсказывать интересующие свойства низкоразмерных материалов, например, стехиометрический состав, электронное строение,

влияние дефектов на структуру и свойства и др. Для оптимизации подобных задач, сокращения времени обработки больших массивов данных, автоматизации анализа полученных результатов и ряда других операций могут быть задействованы современные инструменты *Big Data, Machine Learning, Neural Networks* и *Advanced Analytics.* Использование таких подходов в области материаловедения предоставляет исследователям ряд новых возможностей, в частности, генерации новых экспериментальных моделей наноструктур с предиктивным анализом их электронных свойств, значительно сокращая временные затраты на получение результата по сравнению с традиционными «ручными» режимами моделирования, в которых значительная часть времени может быть затрачена на построение экспериментальной структурной модели и выбор и подготовку входных данных для вычислений. В данной работе предложен подход для оптимизации решения специальной исследовательской задачи – анализа структуры и свойств выбранного материала, который может быть использован для изучения свойств объемных и сверхтонких наноструктур из практически любых неаморфных материалов.

Постановка задачи. Эффективность вычислительных экспериментов для исследования и определения структур и электронных свойств наноматериалов в значительной мере обусловлена качеством программной реализации и используемыми вычислительными моделями компьютерного эксперимента. Хорошо зарекомендовали себя, например, такие общедоступные программные комплексы как *VESTA* (*Visualisation for Electronic and Structural Analysis*) и *OpenMX* (*Open source package for Material eXplorer*) [1–3]. Первый применяется для построения пространственных структур и визуализации их электронных свойств, второй – для определения структурных и электронных свойств материалов с использованием теории функционала плотности, теории псевдопотенциала и базиса численных атомно-центрированных орбиталей. С применением результатов моделирования в *OpenMX* можно оценить возможность использования новых или сложных в получении материалов в наноэлектронике и других приложениях, что сокращает затраты и время на проведение реальных физических экспериментов. Практическая трудность использования компьютерного эксперимента, например, в случае полупроводниковых материалов, обусловлена широким разнообразием параметров, характеризующих их морфологические свойства (такие как симметрия кристаллической ячейки, взаимное расположение кристаллографических плоскостей и пространственная ориентация интерфейса в случае двумерных материалов, взаимодействие и деформация слоев для слоистых структур, разнообразие возможных дефектов, особенно, для композитных составов, и др.). Современные вычислительные средства позволяют достаточно быстро проводить масштабные моделирования наноструктур с помощью программных комплексов, однако, подготовка входных данных для вычислений и описание всех возможных вариаций для моделирования в ручном режиме может занимать значительное время, превышающее многократно затраченное на сам вычислительный эксперимент. Для решения описанных проблем требуются более углубленные, комплексные подходы, сочетающие в себе возможности оптимизации и ускорения процессов сбора данных, моделирования и анализа с помощью инструментов *Big Data, Machine Learning & Generative AI.*

Авторами данной работы для оценки и анализа результатов рентгенодифракционного анализа материалов с возможностью прогнозирования свойств наноструктур ранее было применено предиктивное машинное обучение [4, 5]. Данный подход включал в себя создание нейронной сети и её обучение с помощью эталонных образцов дифрактограмм наноструктур с описанными свойствами. Такая нейросеть смогла бы предложить вероятностное определение их характеристик с определенной степенью точности, зависящей от количества циклов обучения и степени подробности обучающего материала, в случае запроса. Однако, для прогнозирования электронных свойств реализация данного подхода оказалась весьма трудоемкой, что обусловлено наличием

неочевидных зависимостей электронных свойств от структурной конфигурации в материалах с понижением размерности [6].

Подход, основанный на использовании больших языковых моделей (*LLM – Large Language Models*) с применением генерации дополненного поиска (*RAG – Retrieval-Augmented Generation*) ранее применялся авторами для попыток экспресс-анализа больших объемов данных для определения ключевых характеристик наноструктур, опубликованных в открытых источниках. Была реализована специальная методология “обогащения” контекста в виде наполнения специализированных баз знаний (*AI Knowledge Bases*) и, соответственно, расширения области знаний фундаментальной модели (в данном решении используется фундаментальная модель *Anthropic Claude 3.5 Sonnet*). Затем при обращении пользовательского запроса в модель происходила сверка с *RAG*-контекстом и, с учетом специализированного контекста из векторной базы данных, осуществлялась генерация ответа пользователю, при этом релевантный поиск по сотням гигабайт векторных данных языковая модель занимал несколько секунд. При использовании описанного подхода пользователь получает ответ в виде наиболее вероятных словосочетаний и предложений, которые описывают интересующие характеристики наноструктур [7].

Методология и архитектура. Ряд преимуществ описанных подходов, тем не менее, не может решить одного из трудоемких вопросов, стоящих перед исследователями в области материаловедения, заключающегося в поиске наиболее вероятно подходящих параметров кристаллической структуры материала, комбинация которых обеспечит, во-первых, сходимость расчета, а во-вторых требуемое значение ширины запрещенной зоны исследуемого материала. В рамках данной работы реализуется комбинированный подход, использующий *Generative AI* / *LLM* для накопления, структурирования и классификации открытых данных о наноструктурах, а также классическое машинное обучение (*Discriminative ML*), основанное на известных классифицирующих алгоритмах (*XGBoost, LightGBM, Convolutional Neural Network, LTSM – Long-Term Short Memory*). Следует отметить, что применяется продвинутый подход *LLM Agentic AI* (т. наз. агент), который, в свою очередь, основан на использовании описанной выше фундаментальной модели *LLM* вместе с обогащением *RAG*. При этом агент позволяет создавать сложные сценарии поведения чат-бота, хранение и анализ истории диалогов с пользователем, автоматизацию программных действий, таких, как вызовы *REST API* методов, сохранение данных в *SQL*, трансформация собранных данных в виде таблиц, и др. Эти действия описываются с помощью *NLP* (*Natural Language Processing*), фактически, «человеческим языком», вместо написания сложного программного кода, традиционно основанного на циклах, условном выполнении, и т.д. (рисунок 1).

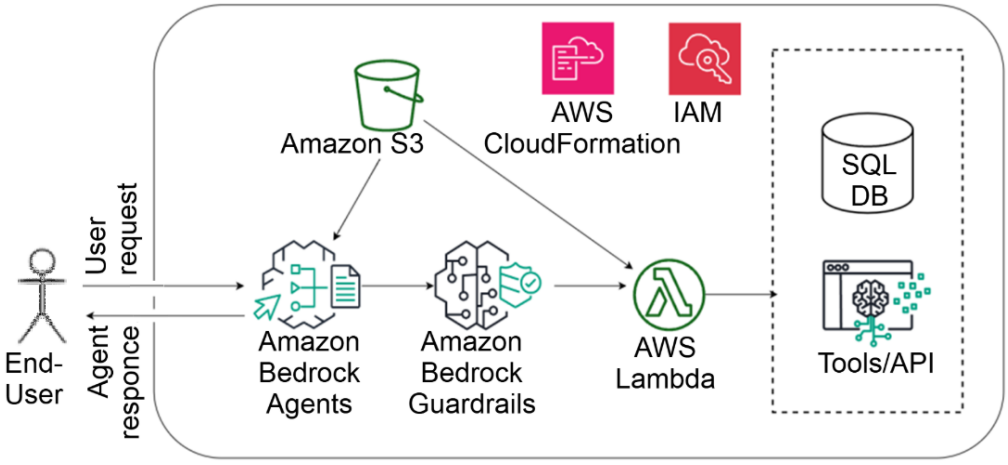

Рисунок 1. Архитектура реализации *Agentic AI* на основе облачного сервиса *AWS* с использованием *Bedrock Agents* и *AWS Lambda functions*

В рамках данной работы также разработан пользовательский интерфейс, который позволяет с помощью *LLM*-инструкций (промпт-инжиниринг) и агентов пополнять базу знаний в виде текстовых данных произвольного формата, подходящих под заданную тематику изучения свойств наноструктур. С помощью интерфейса реализуется возможность вести диалог с *AI*-чат-ботом, производить выборки данных по интересующим материалам с возможностью задания нужных критериев (симметрия кристаллической решетки, стехиометрический состав, ширина запрещенной зоны и др.), а также сохранять их в структурированном виде для последующего развития предиктивной нейронной сети.

Описываемая нейронная сеть была создана на основе облачного сервиса *AWS SageMaker*, который является полноценной средой разработки нейронных сетей и моделей с широкими встроенными возможностями по обработке и классификации тренировочных данных (*SageMaker Data Wrangler*), и автоматизацией подготовки, обучения и разворачивания *ML* моделей для вероятностных классификаций (*SageMaker Pipelines*). В представляемом решении данные, получаемые с помощью *Agentic AI* в виде *csv*-файлов проходят обработку в сервисе *SageMaker Data Wrangler*, которая включает классификацию исходных данных, фильтрацию заведомо непригодных данных, обогащение и трансформацию (*k-nearest neighbors, dropout, oversampling*). Подготовленные данные позволяют далее запускать тренировочные процессы для нейронной сети. На рисунке 2 представлена обобщенная схема работы созданной нейронной сети, предоставленная сервисом *AWS SageMaker*.

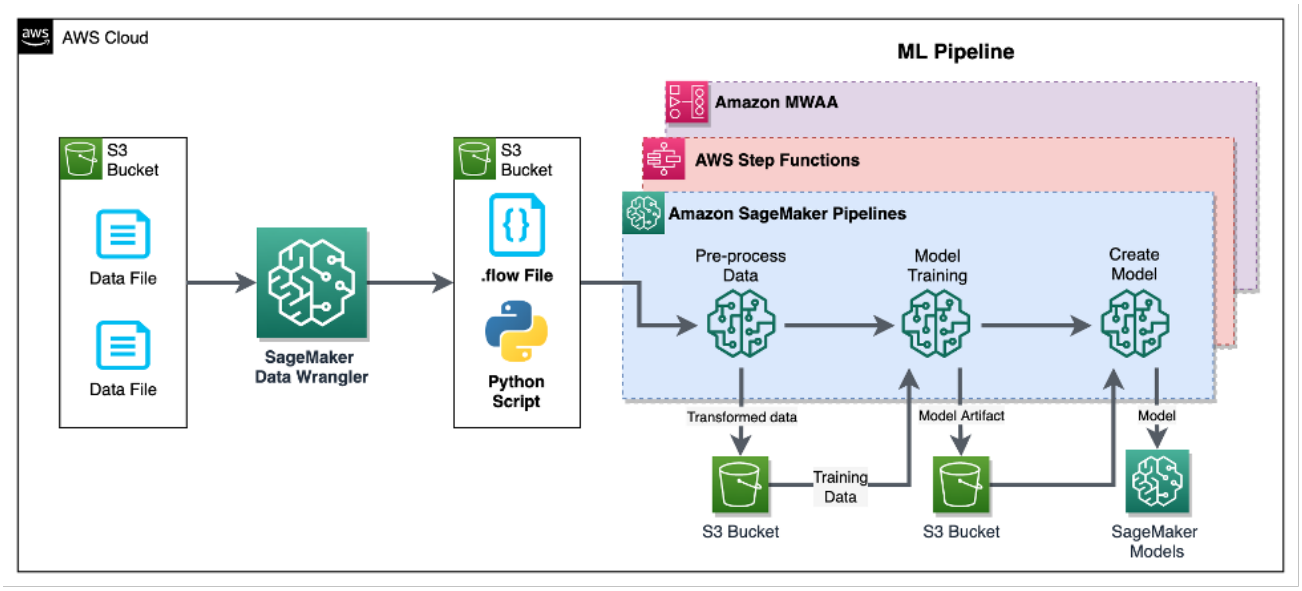

Рисунок 2. Автоматизация подготовки и обработки данных, тренировки нейронной сети и доставки предиктивного сервиса с помощью *SageMaker Data Wrangler* и *SageMaker Pipelines* (скриншот рабочего окна конструктора сервиса)

Следует отметить, что классификация атрибутов нейронной модели в виде точных чисел зачастую демонстрирует плохую сходимость (*ML convergence*) и, как следствие, низкую точность предсказаний свойств исследуемых материалов. В данной работе для нивелирования этой проблемы применяли специальный подход группирования численных значений по диапазонам (*numerical binning*), который на порядки уменьшает количество различных уникальных атрибутов модели за счет группирования близких по значению численных свойств. Такой метод позволяет значительно улучшить сходимость тренировки нейронной сети и качество ее предсказаний, но в то же время может быть ухудшена либо потеряна связь диапазона с исходным точным значением численного параметра вследствие нелинейных сложных зависимостей свойств внутри конкретного диапазона чисел (например, с шириной запрещенной зоны), что в свою очередь, также потребует корректировки.

Разработанное программное решение (использован язык *Python*) способно создавать большие массивы файлов с псевдо-описанием различных возможных конфигураций наноструктур, представляющих потенциальный научный интерес. Данные файлы используются в качестве входных данных программного комплекса *OpenMX*, который на их основе определяет существенные свойства материала (в данном исследовании – ширина запрещенной зоны). Поскольку реальное время расчета тысяч всевозможных комбинаций наноструктур с использованием *OpenMX* занимает значительный период времени, данные массивы потенциально интересных конфигураций предварительно проходят предиктивную фильтрацию в сервисе *SageMaker Inference*, который на основе обученной нейронной модели классифицирует входные данные и отсеивает, по его мнению, заведомо непригодные для расчета. Таким образом выборка конфигураций, перспективных для непосредственного расчета, может быть сокращена в десятки раз. Разработанный подход был проверен на совокупности наноструктур на основе широкозонного сегнетоэлектрика титаната бария ($BaTiO_3$). Выборочные контрольные расчеты отобранных сервисом структур показали, что в подавляющем количестве случаев классификация подходящих и неподходящих структур совпадает с диапазонами реальных расчетов ширины запрещенной зоны в пакете *OpenMX*. Таким образом, предварительная выборка способна значительно ускорить проведение подобных исследований и сократить время реального потребления вычислительных ресурсов.

Заключение. Прогрессивные вычислительные подходы могут и должны внести значительный вклад в развитие современного материаловедения. Применение нейронных сетей (*Generative AI, LLM*) в качестве больших языковых моделей для сбора, классификации и обобщения свойств наноструктур и наноматериалов в сочетании со статистическим машинным обучением демонстрирует высокий потенциал применения, в особенности, для повышения уровня автоматизации ряда исследовательских задач при значительном сокращении временных и трудовых затрат на проведение вычислительных экспериментов. Предложенный в данной работе подход предусматривает непрерывное дополнение и развитие базы знаний *LLM* модели в рамках специализированного научного контекста (в рассматриваемом случае – в области наноматериаловедения), а также развитие предиктивной нейронной сети для предварительного статистического анализа моделируемых наноматериалов, что позволит значительно снизить временные и аппаратные затраты на моделирование с помощью автоматизированной классификации перспективных конфигураций наноструктур. Данный подход при подборе соответствующей базы знаний может быть расширен также для предиктивного анализа свойств более сложных наноструктур, в частности, бислоев, Ван-дер-Ваальсовских структур, гетеросистем различного назначения, сильнолегированных материалов и т.д.

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] 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 [Silhavy, R., Silhavy, P., Prokopova, Z. (eds)] / Springer Cham. 2019. Vol 1047. P.208–215.

Авторский вклад

Шиманский Никита Андреевич – концептуализация и разработка решения, описание принципов работы агента, работа с облачными сервисами, написание и адаптация программного кода, тестирование решения, формирование структуры статьи.

Баглов Алексей Викторович – постановка задачи исследования, описание принципа работы решения, анализ и верификация результатов, тестирование решения, обеспечение проверки результатов в пакете *OpenMX*.

Хорошко Людмила Сергеевна – постановка задачи исследования, концептуализация решения, определение структуры исследования, тестирование решения, анализ и верификация результатов, формирование структуры статьи, общее руководство и менеджмент проекта.

STATISTICAL ANALYSIS AND MODELING OF NANOMATERIALS PROPERTIES USING LARGE LANGUAGE MODELS, AGENTIC AI & MACHINE LEARNING

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

A.V. Baglov
Senior researcher in the R&D Lab of Energy-effective materials and technologies, Faculty of Physics, BSU

L.S. Khoroshko
Lead researcher in the R&D Lab of Energy-effective materials and technologies, Faculty of Physics, BSU

Abstract. This paper investigates the possibility of the combination of generative machine learning and classical neural networks as methods for working with Big Data for the predictive analysis of the electronic properties of semiconductor nanomaterials, without limiting the generality of this approach to other crystalline materials and heterostructures.

Keywords: machine learning, neural networks, inference, large language models, agentic ai, big data, nanomaterials, predictive analysis.