УДК 004.65+004.85

ЭКСПРЕСС-АНАЛИЗ СТРУКТУРНЫХ И ЭЛЕКТРОННЫХ СВОЙСТВ НАНОМАТЕРИАЛОВ МЕТОДАМИ BIG DATA, LARGE LANGUAGE MODELS & GENERATIVE AI

Шиманский Н.А.[1,2], Баглов А.В.[2,3], Хорошко Л.С.[2,3]

[1]ООО «АндерсенБел» (ПВТ), Минск, Республика Беларусь, nikita.shymanski@gmail.com

[2]Белорусский государственный университет, Минск, Республика Беларусь

[3]Белорусский государственный университет информатики и радиоэлектроники, Минск, Республика Беларусь

Аннотация: В данной работе рассматривается возможность использования генеративного машинного обучения и методов работы с большими данными для предиктивного анализа электронных свойств сверхтонких наноструктур на основе полупроводниковых материалов, не ограничивая при этом общность данного подхода для иных кристаллических материалов.

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

I. ВВЕДЕНИЕ

Успешное развитие науки в области синтеза и исследования современных и перспективных наноматериалов на сегодняшний день практически невозможно без использования вычислительных средств и компьютерной автоматизации. Стремительное развитие нейронных сетей и программных интерфейсов для них (Machine Learning, 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 исследуемых материалов, например, стехиометрический состав, электронные свойства и влияние на них дефектов и др. Современные инструменты BigData и Advanced Analytics могут быть привлечены для оптимизации подобных задач, сокращении времени обработки больших массивов данных, автоматизации анализа полученных результатов и т.д. В частности, в области материаловедения такие подходы предоставляют исследователям возможности генерации новых экспериментальных моделей наноструктур с предиктивным анализом их электронных свойств фактически в режиме реального времени, в отличие от «ручных» режимов моделирования. В данной работе рассмотрен пример решения специальной исследовательской задачи – моделирования и анализа структуры и свойств выбранного полупроводникового материала, который может быть использован также для изучения свойств объемных и сверхтонких наноструктур из практически любых неаморфных материалов.

II. АНАЛИЗ И 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 моделирования. 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 используется для моделирования пространственных структур и визуализации их электронных свойств, второй – для определения электронных свойств (зонная структур и т.д.). на основе теорий функционала плотности (DFT), нормосохраняющих псевдопотенциалов и псевдоатомных локализованных базисных функций. Результаты моделирования в OpenMX позволяют учесть электронные свойства исследуемых материалов для анализа возможности использования в полупроводниковой наноэлектронике, что востребовано для проведения предварительных вычислительных экспериментов с новыми материалами. Практическая трудность реализации данного подхода обусловлена широким разнообразием параметров полупроводниковых структур, характеризующих их морфологические свойства (такие как симметрия кристаллической ячейки, взаимное расположение кристаллографических плоскостей и пространственная ориентация интерфейса в случае двумерных материалов, деформация и взаимодействие слоев для слоистых структур и др.). Аккуратное описание всех этих параметров требует значительных временных затрат и осуществляется на этапе формирования входного файла, задающего параметры вычислений, преимущественно в ручном режиме. Процесс моделирования в различных программных пакетах в совокупности с подготовкой исходных данных может занимать значительное время, при этом

прогнозная или вероятностная оценка результата проводимого исследования практически невозможна.

Для решения описанных проблем требуется комплексный подход, сочетающий в себе возможности оптимизации и ускорения процессов моделирования и анализа наноструктур с помощью инструментов Big Data & Machine Learning. Авторы данной работы уже применяли подобные подходы для оценки и анализа результатов одного из базовых методов исследований свойств и структуры кристаллических материалов – дифракции рентгеновских лучей [4, 5]. Предиктивное машинное обучение было применено для предварительного прогнозирования свойств наноструктур и включало в себя создание нейронной сети и её обучение с помощью эталонных образцов дифрактограмм наноструктур с описанными свойствами. В случае появления задачи исследования структурных свойств кристаллических материалов такая нейросеть смогла бы предложить вероятностное определение их характеристик с определенной степенью точности, зависящей от количества циклов обучения и степени подробности обучающего материала. Однако, реализация данного подхода для прогнозирования электронных свойств весьма трудоемка, что обусловлено появлением неочевидных зависимостей электронных свойств от структурной конфигурации в материалах с понижением размерности, а также, как уже упоминалось ранее, большим количеством варьируемых параметров для двумерных и сверхтонких (до 10 монослоев) наноструктур. Обработка таких зависимостей с использованием универсального алгоритма является проблематичной, а порой и невозможной, в результате чего вероятностный (предиктивный) прогноз свойств с применением нейросети снижается вплоть до получения недостоверного результата [6].

III. ОПИСАНИЕ РЕШЕНИЯ

Использование больших языковых моделей (Generative AI, LLM – Large Language Models) для решения описанной проблемы может быть использовано как принципиально новый подход к использованию нейронных сетей и машинного обучения в физике и материаловедении наноструктур [4]. Для использования LLM в специализированных областях предлагается специальная методология “обогащения” контекста и, соответственно, расширения области знаний нейронной сети, которая носит название генерации дополненного поиска (Retrieval-Augmented Generation, RAG). Данный метод представляет собой оптимизацию выходных данных большой языковой модели, в результате которой она ссылается на “обогащенную” (дополненную) базу знаний, т.е. перед генерацией ответа выходит за пределы своих источников обучающих данных (рис. 1). RAG дополнительно расширяет возможности LLM на определенные домены или внутреннюю базу данных организации, при этом переобучения модели не требуется. Примером быстроразвивающихся языковых моделей с поддержкой RAG на сегодняшний день могут являться Meta Llama 3, Anthropic Claude, Amazon Titan и ряд других менее популярных у широкого круга специалистов.

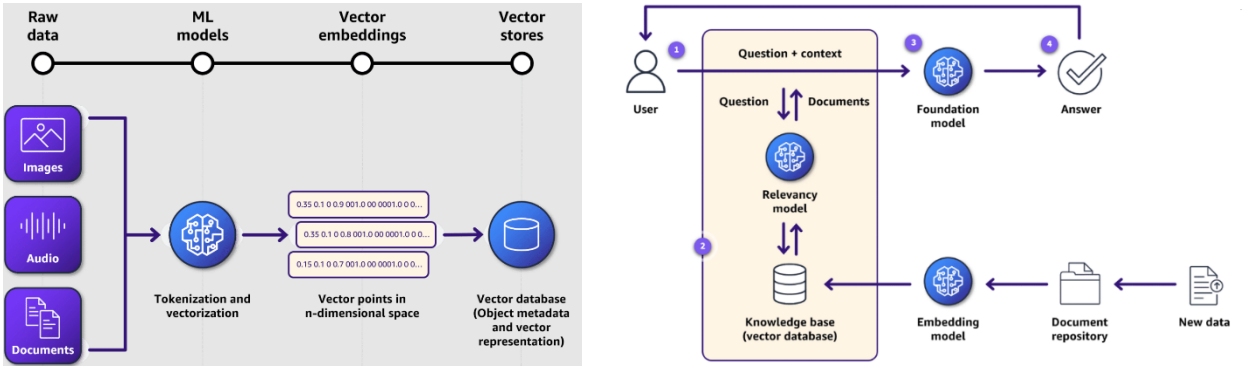

Рисунок 1. Схема процессов привлечения векторных баз данных для обогащения ответов языковой модели при пользовательском запросе в LLM

В рамках решения описанной задачи RAG-обогащение происходит в несколько этапов. На начальном этапе накапливается релевантная информация в текстовом и бинарном формате (форматы .txt, .doc(x), .xls(x), .pdf и др.), имеющем произвольную организацию и структуру, но содержащем определенные количественные описания свойств наноструктур и их качественную интерпретацию. Примерами таких источников текста могут служить научные публикации и монографии, отчеты о лабораторных исследованиях и измерениях, данные по моделированию определенных структур, обзорные публикации, рефераты и выводы на их основе, теоретические исследования фундаментальных свойств материалов и т.д. Также с использованием RAG можно работать с графическими материалами, сканировать и распознавать изображения, т.е. языковая модель может обучаться с использованием

графиков зависимостей, изображений структур, визуализации экспериментальных данных и др. Одним из важных преимуществ использования такой методики является отсутствие строгих требований к оглавлению и структурированию содержания материала, а также, фактически, ограничений по размерам обрабатываемых файлов. По мере сбора и накопления обучающего контента подключается специализированная векторная база данных (например LanceDB, AWS OpenSearch), которая трансформирует этот контент в бинарный вид и интегрируется с языковой моделью (рис. 2).

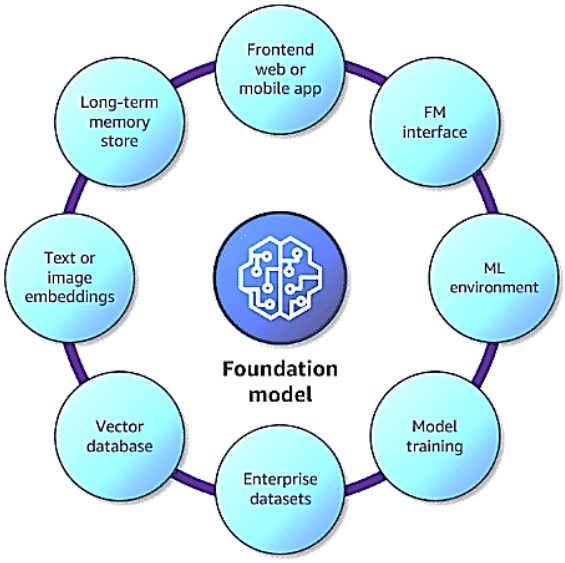

Рисунок 2. Архитектурная диаграмма в виде программного интерфейса с пользовательским вводом для описанного решения

Затем при обращении пользовательского запроса в модель происходит сверка с RAG-контекстом и с учетом специализированного контекста из векторной базы данных осуществляется генерация ответа пользователю, при этом релевантный поиск по десяткам и сотням терабайт векторных данных языковая модель осуществляет за несколько секунд, а пользователь получает ответ в виде наиболее вероятных словосочетаний и предложений. При этом в отличие от традиционных нейронных сетей, для реализации данного подхода не требуется дополнительных тренировок и переобучений языковой модели, что сокращает вычислительные и временные затраты на использование искусственного интеллекта.

Описываемый подход предлагается к реализации в виде кроссплатформенного программного решения, которое находится в стадии разработки и накопления векторных баз данных. Перспективность развития данной идеи подтверждает ряд предварительных экспериментов в виде диалогов (prompting) с искусственным интеллектом. В частности, используемый для взаимодействия RAG чат-бот в состоянии определить характеристики зонной структуры (ширина запрещенной зоны, энергия Ферми) в наноструктуре на примере широкозонного сегнетоэлектрика титаната бария ($BaTiO_3$). Используя “обогащенный” контекст и созданную заранее векторную базу данных, чат-бот определил эти значения и сгенерировал корректный ответ на запрос, при этом тестируемая языковая модель не обладала запрашиваемыми параметрами в явном виде, но смогла провести статистическое сопоставление и сгенерировать верный ответ. При этом прогностические запросы по инжинирингу запрещенной зоны (изменение ее ширины путем модификации кристаллической решетки или внедрения допантов и дефектов) чат-бот не смог корректно обработать: вместо этого языковая модель предлагала другие химические соединения с искомыми параметрами запрещенной зоны. Это показывает важность создания баз данных достаточной полноты и их своевременной актуализации, из чего следует важность наличия открытых результатов исследований для глобализации научной сети и полноценного внедрения современных технологий в наукоемкие сферы по всему миру.

IV. ЗАКЛЮЧЕНИЕ

Для развития современного материаловедения характерны, в общем случае, две противоположных тенденции. С одной стороны, соревновательный характер ведения исследований между научными группами разных стран, что предполагает закрытый формат результатов наряду с публикованием в изданиях, имеющих ограниченный круг доступа, что в сочетании со своевременным патентованием и оформлением ноу-хау помогает сохранить эксклюзивность результатов исследований. С другой стороны, неизбежная глобализация всех процессов научных исследований привела к созданию

мощнейших мировых исследовательских коллабораций и научных центров (например, Объединенный институт ядерных исследований в г. Дубне, РФ и др.), в которых совместное получение новых знаний и использование результатов является основой существования проектов и устойчивого развития всей исследовательской инфраструктуры. Использование современных информационных технологий может стать дополнительным средством обеспечения эффективного взаимодействия между учеными и исследователями всех стран, в том числе, для обмена актуальными результатами, сопоставления и верификации новых данных, облегчения поисковой исследовательской работы и повышения ее эффективности. Применение нейронных сетей (Generative AI, LLM) в качестве больших языковых моделей для исследования и анализа свойств наноструктур является перспективным, особенно, для повышения уровня автоматизации ряда исследовательских задач при значительном сокращении временных и трудовых затраты на обработку. Предложенный в данной работе подход предусматривает постоянное развитие и обучение LLM модели в рамках специализированного научного контекста (в рассматриваемом случае - в области наноматериалов), что будет способствовать по мере накопления заданного контекста увеличению точности прогнозного анализа и способности искусственного интеллект не только корректно производить поиск по имеющимся данным, но и предлагать новые наноструктуры с требуемыми и предсказанными свойствами.

БЛАГОДАРНОСТЬ

Исследования частично поддержаны в рамках НИР 4 по заданию № 2.25 ГПНИ «Материаловедение, новые материалы и технологии».

ЛИТЕРАТУРА

[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.

EXPRESS ANALYSIS OF THE STRUCTURAL AND ELECTRONIC PROPERTIES OF NANOMATERIALS USING BIG DATA, LARGE LANGUAGE MODELS & GENERATIVE AI

N. Shimansky[1,2], A. Baglov[2,3], L. Khoroshko[2,3]

[1]AndersenBel company (HTP), Minsk, Republic of Belarus, nikita.shymanski@gmail.com

[2]Belarusian State University, Minsk, Republic of Belarus

[3]Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Abstract: This paper explores the possibility of using generative Machine Learning and Big Data methods for predictive analysis of the electronic properties of ultrathin and nanostructures based on crystalline semiconductor materials, without limiting the generality of this approach to other crystalline materials.

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