УДК 004.457

АВТОМАТИЗАЦИЯ ОБРАБОТКИ AB INITIO РАСЧЕТОВ ЗОННОЙ СТРУКТУРЫ ДЛЯ ПОДГОТОВКИ ДАТАСЕТОВ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Баглов А.В.[1,2], Хорошко Л.С.[1,2]

[1]Белорусский государственный университет, Минск, Республика Беларусь, baglov@bsu.by

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

Аннотация: описаны новые функциональные возможности кроссплатформенного анализатора зонной структуры материалов для пакетов OpenMX и CASTEP. Благодаря переходу к модульной архитектуре «Бэкенд – Ядро – Фронтенд» обеспечена высокая производительность и масштабируемость системы. Продемонстрировано применение инструмента для автоматизированной подготовки физически обоснованных наборов данных, адаптированных для обучения моделей машинного обучения и больших языковых моделей. Гибкая система бэкендов позволяет оперативно расширять перечень поддерживаемых научных пакетов.

Ключевые слова: анализ данных, постобработка, C++, gnuplot, OpenMX, моделирование, электронная структура, большие языковые модели.

I. ВВЕДЕНИЕ

В наноматериаловедении для расчёта структурных, электронных и оптических свойств применяются как классические, так и квантово-механические подходы. Наибольшую точность обеспечивают первопринципные (*ab initio*) методы [1], однако их вычислительная стоимость ограничивает размер исследуемых систем ~100 атомов. При дизайне новых материалов, требующем перебора множества составов и структур, прямой расчёт становится нецелесообразным. Альтернативой служит использование больших языковых моделей (БЯМ), способных классифицировать материалы и предсказывать их свойства без проведения расчётов, эффективно сужая пространство поиска. В отличие от графовых нейронных сетей, БЯМ непосредственно обрабатывают стандартизированные текстовые представления структур (CIF, POSCAR), выявляя закономерности «структура–свойство» через механизмы внимания.

Для обучения БЯМ необходим датасет, включающий кристаллическую структуру, химический состав и зонную структуру. Экспериментальная сложность определения последней обусловливает использование *ab initio* расчётов для генерации данных, при этом рост числа электронных состояний с размером системы увеличивает затраты на подготовку выборок. Полученные данные требуют существенной постобработки. Авторы используют пакет OpenMX [2–9] и ранее разработали для него анализатор зонной структуры [10]. В данной работе представлено расширение его функциональности и продемонстрирована эффективность в качестве связующего звена (*middleware*) для автоматизированной генерации обучающих выборок для БЯМ.

II. ОПИСАНИЕ АНАЛИЗАТОРА ЗОННОЙ СТРУКТУРЫ

Взаимодействие с анализатором, как и с пакетом OpenMX, осуществляется через механизм CLI (Command Line Interface – интерфейс командной строки). Приложение полностью написано на языке C++ и не имеет никаких зависимостей, кроме математической библиотеки и рантайма C++, что обеспечивает высокую производительность и переносимость кода. Сборка осуществляется с помощью утилит make и nmake под Linux и Windows соответственно. Исходные «сырые» данные разбиваются анализатором на две части: свойства решетки и зонная структура. Результаты работы анализатора могут направляться в стандартный поток вывода, текстовые файлы или комбинированным образом – и на экран, и в файл. Для вывода ошибок используется стандартный поток ошибок. Текущая версия 0.4.2.0 значительно доработана в сравнении с ранее описанной версией 0.3.4.2 [10]. Пересмотрено и расширено представление результатов анализа. Помимо изменений в представлении результатов анализа проведен значительный рефакторинг исходного кода с переходом к модульной архитектуре «Бэкенд – Ядро – Фронтенд». Бэкенд занимается чтением сырых данных, конвертаций и приведением типов для заполнения специальной структуру данных, передаваемой в ядро анализатора. Такой подход позволяет отделить функциональные возможности ядра от формата входных данных, что позволило нам подготовить бэкенд для широко известного пакета квантово-механического моделирование CASTEP в дополнение к основному бэкенду для OpenMX. В ядре анализатора также сделан ряд изменений, облегчающих обработку и вывод данных за счет ряда оптимизаций и

использования шаблонных функций. Фронтенд отвечает за направление данных либо в стандартный поток вывода, либо в файлы. Поскольку используется семантика перемещения и RVO (Return Value Optimization), то переход к модульной архитектуре не привнес дополнительных расходов, а оптимизация позволила добиться некоторого ускорения в сравнении с ранней версией. В версии 0.4.2.0 была переработана логика взаимодействия с анализатором. Ниже приведен список ключевых изменений и новых параметров:

-kpts – новый параметр. Заменяет -lk. Выводит нумерованный список k-точек;

-l – выводит информацию о прямой и обратной решетках;

-lk – функционал изменен. Дополнительно выводит информацию о сегментах обхода обратной решетки вдоль особых точек высокой симметрии;

-exact – новый параметр. Активирует вывод значений повышенной точности в блоке SUMMARY;

-vbm – удален. Функциональность перекрывается параметром -exact;

Полностью переработана и расширена поддержка пакета научной визуализации gnuplot в Linux и Windows. Поддерживается работа как с OpenMX, так и с CASTEP. Для работы с gnuplot используются параметры -g и -gp. Задание параметра -g приводит к созданию исполняемого файла (скрипта) с расширением cmd в Windows и расширением plt в Linux. Запуск скрипта по двойному клику инициализирует gnuplot в интерактивном режиме для просмотра зонной структуры. Задание параметра -gp заменяет интерактивную работу gnuplot генерацией изображения в формате png. Настройки вывода (шрифты, толщина линий, сетка) по умолчанию соответствуют стандартам научных публикаций (рис. 1). Файлы cmd/plt имеют формат UTF-8 и при необходимости могут быть открыты в текстовом редакторе и скорректированы под требования пользователя, что позволяет использовать спецсимволы в подписях при необходимости.

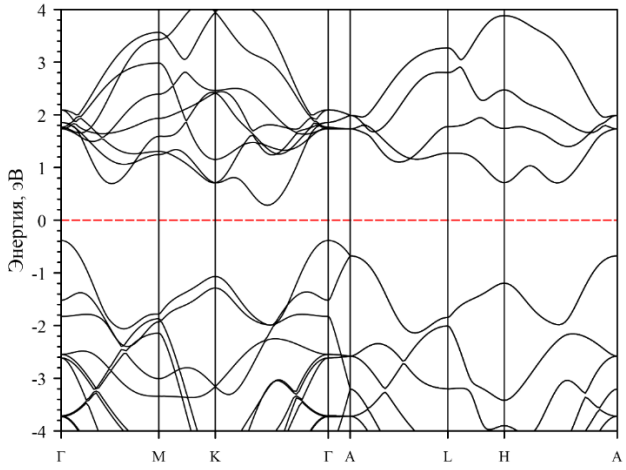


Рисунок 1. Зонная структура 2H-MoS_2, построенная с помощью сгенерированного скрипта для gnuplot. Вручную скорректирована область энергий и соотношение сторон

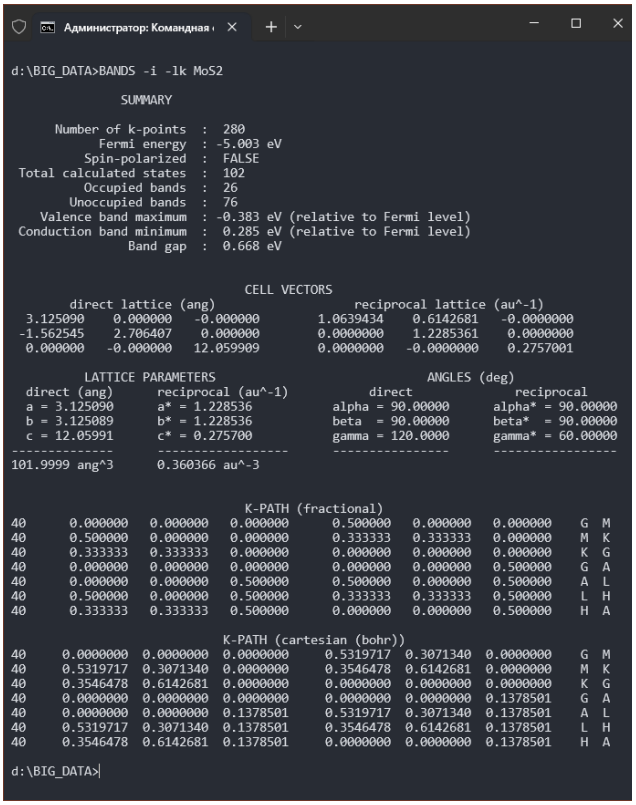


Рисунок 2. Пример результата работы анализатора в операционной системе Windows на тестовой задаче

III. ИСПОЛЬЗОВАНИЕ В КАЧЕСТВЕ СВЯЗУЮЩЕГО ЗВЕНА

Анализатор может функционировать в режиме предоставления описания исследуемой системы, что делает его эффективным инструментом для формирования и оптимизации датасетов.

В блоке SUMMARY дается краткая информация о электронной подсистеме. Важно отметить, что значение энергии Ферми, получаемое в расчетах, специфично для каждой конкретной реализации численного метода (конструкция псевдопотенциалов, вид базис, алгоритм смешивания электронной плотности и т.д.). По этой причине оно не может напрямую сравниваться между различными программными пакетами и, тем более, экспериментальными данными. Поскольку анализатор это CLI-приложение, можно подготовить скриптовый препроцессор для автоматизированного извлечения необходимых данных и формирования датасета. В данном контексте представляют интерес следующие параметры: число k-точек, наличие или отсутствие спиновой поляризации, число занятых состояний, положения экстремумов зон и ширина энергетической щели. При подготовке данных о кристаллической структуре можно прочитать входной файл для OpenMX, где содержится информация о векторах трансляции прямой решетки и позициях атомов в ней. Более полную информацию о свойствах решетки также можно найти в выходных данных анализатора, блок CELL VECTORS. Для удобства восприятия мы выбрали дифференцированное представление параметров прямой решетки в ангстремах, а обратной решетки в атомных единицах (Борах), дополнительно предоставляется информация об объемах прямой и обратной решеток, а также углах в них. В контексте подготовки датасета можно использовать данные либо прямой решетки, либо обратной, поскольку они определяются взаимно однозначно [11]. Блок данных, связанный с особыми точками высокой симметрии и промежуточными между ними информативен в первую очередь для исследователя и служит для верификации корректности выбранного пути в обратной решетке.

Поскольку для исследователя более интересны электронные состояния вблизи потолка валентной зоны и дна зоны проводимости, анализатор позволяет извлечь зоны в заданном диапазоне индексов. Выбор оптимального числа зон для обучения модели не всегда очевиден, поэтому мы предлагаем опираться на предоставляемую анализатором информацию о глобальных минимумах и максимумах каждой зоны. В рассматриваемом примере (см. рисунки 1 и 2) зоны с первой по двенадцатую лежат на ≈ 14–60 эВ ниже потолка валентной зоны, а потому могут быть отброшены. Таким образом остается 14 валентных зон, которые необходимо извлечь. Аналогично можно поступить и с зонами проводимости. В принципе, можно еще сильнее ограничить область интересующих энергий, например ±5 эВ, что уменьшит число зон до 20 (12 валентных и 8 проводимости), что пятикратно уменьшает объем данных для датасета. Такая оптимизация радикально снижает вычислительные затраты на обучение БЯМ, сохраняя при этом всю физически значимую информацию о поведении электронов вблизи запрещенной зоны.

III. ЗАКЛЮЧЕНИЕ

Представлено развитие функциональных возможностей кроссплатформенного анализатора зонной структуры. Показано, что переход к модульной архитектуре «Бэкенд – Ядро – Фронтенд» и использование возможностей стандарта C++ и выше (семантика перемещения, RVO) позволили повысить высокую производительность при нулевых внешних зависимостях. Инструмент полностью кроссплатформенный и легко интегрируется в автоматизированные конвейеры обработки данных. Добавлен бэкенд для широко известного пакета для квантово-механических расчетов CASTEP. Переработана система интеграции с gnuplot, обеспечивающая генерацию изображений журнального качества в автоматическом режиме. Поддержка кодировки UTF-8 позволяет корректно отображать кириллицу и специальные символы, упрощая подготовку научных публикаций. Предложенный инструмент решает критическую задачу подготовки качественных, физически обоснованных обучающих выборок. Дифференцированное представление параметров решетки и извлечение основных характеристик электронной подсистемы позволяют минимизировать смещение моделей при обучении на подготовленных данных с сохранением физически значимой информации о состояниях вблизи уровня Ферми. Таким образом, разработанный анализатор закладывает фундамент для создания масштабируемых систем предсказательного моделирования, способных значительно ускорить поиск и дизайн новых функциональных наноматериалов, минимизируя необходимость в дорогостоящих квантово-механических расчетах.

БЛАГОДАРНОСТЬ

Работа частично поддержана БРФФИ (проект № Т25МЭ-007), а также в рамках задания № 1.2.2 «Тонкопленочные и двумерные полупроводниковые материалы на основе углерода, халькогенидов, в том числе тугоплавких металлов, и неорганических перовскитов для приложений сенсорики, фотовольтаики, нано- и оптоэлектроники» подпрограммы «Физика конденсированного состояния, разработка технологий создания функциональных инструментальных материалов и устройств», НИР «Низкотемпературный синтез и свойства наноструктурированных композитов на основе перовскитных соединений с кластерами редкоземельных элементов и углерода для конвертеров оптического излучения и фотовольтаики» подпрограммы «Наноструктуры, наноматериалы и технологии» ГПНИ

«Современное материаловедение, перспективные материалы и новые технологии»; НИР «Разработка научных основ и методов сенсибилизации, детектирования и фотокаталитической деструкции органических загрязнителей, включая микро- и нанопластики, в водных средах с применением гибридных мультифункциональных нанокомпозитов на основе оксидных полупроводников» подпрограммы «Междисциплинарные исследования и перспективные технологии» ГПНИ «Конвергенция 2030».

Авторы выражают благодарность системному архитектору компании Andersen Lab (ПВТ) Н.А. Шиманскому за полезную дискуссию в области подготовки компактных БЯМ.

ЛИТЕРАТУРА

[1] Shevlin, S. Computational materials design / S. Shevlin, B. Castro, X. Li // Nature Materials. 2021. Vol. 20. P. 727–728. https://doi.org/10.1038/s41563-021-01038-8.

[2] Ozaki, T. Variationally optimized atomic orbitals for large-scale electronic structures / T. Ozaki // Physical Review B. 2003. Vol. 67, No 15. P. 155108. https://doi.org/10.1103/PhysRevB.67.155108.

[3] Ozaki, T. Numerical atomic basis orbitals from H to Kr / T. Ozaki, H. Kino // Physical Review B. 2004. Vol. 69, No 19. P. 195113. https://doi.org/10.1103/PhysRevB.69.195113.

[4] Ozaki, T. Efficient projector expansion for the ab initio LCAO method / T. Ozaki, H. Kino // Physical Review B. 2005. Vol. 72, No 4. P. 045121. https://doi.org/10.1103/PhysRevB.72.045121.

[5] Baglov, A. V. Atom Species Energy Dependence on Magnetic Configurations in the Perovskite Yttrium Orthoferrite / A. V. Baglov, L. S. Khoroshko // Doklady BGUIR. 2021. Vol. 19, No 8. P. 63–67. https://doi.org/10.35596/1729-7648-2021-19-8-63-67.

[6] Baglov, A. V. Crystal Structure and Electronic Properties of Rhenium Disulfide / A. V. Baglov, L. S. Khoroshko // Journal of Applied Spectroscopy. 2022. Vol. 89. P. 860–864. https://doi.org/10.47612/0514-7506-2022-89-5-651-656.

[7] Evolution of Structural and Electronic Properties Standardized Description in Rhenium Disulfide at the Bulk-Monolayer Transition / A. Baglov [et al.] // Heliyon. 2024. Vol. 10. P. e28646. https://doi.org/10.1016/j.heliyon.2024.e28646.

[8] Electronic Structure of Bismuth Ferromanganite $BiFe_{0.5}Mn_{0.5}O_3$ / A. V. Baglov [et al.] // Semiconductors. 2024. Vol. 58, No 13. P. 1054–1059. https://doi.org/10.1134/S1063782624700040

[9] Baglov, A. V. Structural and Electronic Properties of $SmGaGe_2O_7$ Studied by First Principles Methods / A. V. Baglov, L. S. Khoroshko // Inorganic Materials. 2023. Vol. 59, No 1. P. 1–7. https://doi.org/10.1134/S002016852301003X

[10] Баглов, А.В. Кросс-платформенный анализатор зонной структуры материалов для пакета квантово-механического моделирования OpenMX / А.В. Баглов, Л.С. Хорошко // Доклады БГУИР. 2025. Т. 23. No 5. С. 83–92. https://doi.org/10.35596/1729-7648-2025-23-5-83-92.

[11] Ашкрофт, Н. Физика твердого тела: в 2 т / Н. Ашкрофт, Н. Мермин //I–II. М.: Мир. – 1979.

A TOOL FOR AUTOMATING THE PROCESSING OF AB INITIO ZONE STRUCTURE CALCULATIONS FOR PREPARING DATASETS OF LARGE LANGUAGE MODELS

A. Baglov[1,2], L. Khoroshko[1,2]

[1]Belarusian State University, Minsk, Republic of Belarus, baglov@bsu.by

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

Abstract: the new functionality of a cross-platform material band structure analyzer for the OpenMX and CASTEP packages is described. The transition to a modular "backend-core-frontend" architecture ensures high system performance and scalability. This paper shows how to use the tool to automatically prepare physically based datasets for training machine learning and large language models. The flexible backend system enables the rapid expansion of supported scientific packages.

Keywords: data analysis, postprocessing, C++, gnuplot, OpenMX, modeling, electronic structure, large language models.