УДК 51-7

ПРИМЕНЕНИЕ МОЛЕКУЛЯРНЫХ ДЕСКРИПТОРОВ ДЛЯ ПРОГНОЗИРОВАНИЯ СВОЙСТВ ПОЛИМЕРНЫХ И КОМПОЗИЦИОННЫХ МАТЕРИАЛОВ

APPLICATION OF MOLECULAR DESCRIPTORS FOR PREDICTING THE PROPERTIES OF POLYMER AND COMPOSITE MATERIALS

Шайдуллин Нияз Наилевич
магистрант кафедры нефтехимия и химическая технология,
Уфимский государственный нефтяной
технический университет
niazsajdullin@gmail.com

Кичатов Константин Геннадьевич
кандидат химических наук,
доцент кафедры нефтехимия и химическая технология,
Уфимский государственный нефтяной
технический университет
kichatov_k@mail.ru

Просочкина Татьяна Рудольфовна
доктор химических наук,
профессор,
заведующая кафедрой нефтехимия и химическая технология,
Уфимский государственный
нефтяной технический университет
t.r.prosochkina@mail.ru

Shaidullin Niyaz Nailevich
Master of Department of Petrochemistry
and Chemical Technology,
Ufa State Petroleum Technological University
niazsajdullin@gmail.com

Kichatov Konstantin Gennadievich
Candidate of Chemistry Sciences,
Associate Professor of Department
of Petrochemistry and Chemical Technology,
Ufa State Petroleum Technological University
kichatov_k@mail.ru

Prosochkina Tatyana Rudolfovna
Doctor of Chemical Sciences,
Professor, Head of the Department
of Petrochemistry and Chemical Technology,
Ufa State Petroleum Technological University
t.r.prosochkina@mail.ru

Аннотация. Сроки разработки новых полимерных и композиционных материалов при традиционном способе проектирования обычно составляют не менее 10 лет. С целью сокращения времени выхода нового материала на рынок был разработан новый подход, заключающийся в моделировании связи между структурами химических соединений и их свойствами. Для оценки данного подхода была построена модель машинного обучения (МО), обученная на молекулярных дескрипторах.

Ключевые слова: хемоинформатика, машинное обучение, молекулярные дескрипторы, прогнозирование, полимеры.

Annotation. The time required to develop new polymer or composite material using the traditional design method is usually more than 10 years. In order to reduce the time for new material to reach the market, a new approach was developed by modeling the relationship between the structures of chemical compounds and their properties. To evaluate this approach, a machine learning (ML) model trained on molecular descriptors was built.

Keywords: chemoinformatics, machine learning, molecular descriptors, prediction, polymers.

Введение

Полимеры находят свое применение в аэрокосмической, строительной, медицинской, энергетической и пищевой промышленностях [1]. Традиционно синтез и характеристика новых полимеров осуществлялись методом проб и ошибок. Данный подход требует значительных временных и вычислительных ресурсов, а также привлечения квалифицированных специалистов [2]. Высокая сложность и размер химического и структурного пространства возможных соединений значительно ограничивают эффективность такого подхода при поиске материалов с заданными свойствами [3].

В последние годы развитие вычислительных технологий и методов искусственного интеллекта способствовало широкому распространению машинного обучения (МО) в решении задач регрессии и классификации в различных областях науки и техники. МО катализировало развитие материаловедения, применяя компьютерные алгоритмы для расширения возможностей исследования, понимания, моделирования и симуляции различных материалов. Эти алгоритмы, вкупе с творческим потенциалом исследователя, могут внести значительный вклад в открытые новых материалов для технологий будущего [4, 5].

Методы МО позволяют эффективно обрабатывать многомерные данные, выявляя как линейные, так и нелинейные зависимости между структурой и свойствами полимеров. В результате модель МО может быть развернута с высокой точностью с относительно низкими вычислительными затратами.

Сбор данных

Сбор данных является ключевым этапом в проектировании полимеров при помощи МО. Надежность прогнозирования тесно связана с достаточностью и достоверностью данных. Тем не менее, потребность в более релевантных данных является актуальной проблемой из-за ограниченности и стоимости лабораторных исследований. Наиболее распространенными источниками данных являются научные публикации и открытые базы данных [1, 6].

Данные, представленные в научных публикациях, имеют наибольшую достоверность, так как получены в результате лабораторных экспериментов [7]. Однако их извлечение требует значительных усилий. Открытые базы данных обеспечивают доступ к большому количеству информации и инструментам анализа, но могут содержать неточные или синтетические данные, что снижает надежность.

В данной работе используется база данных, сформированная на основе расчетов молекулярной механики и молекулярной динамики, включающая 1067 наблюдений [8]. В датасете представлены SMILES-строки мономеров и вычисленные физико-химические свойства полимеров. Для прогнозирования выбраны плотность (density), показатель преломления

(refractive_index), модуль упругости (bulk_modulus) и сжимаемость (compressibility).

Расчет дескрипторов

Для преобразования химических данных в машиночитаемый формат используются молекулярные дескрипторы, способных количественно характеризовать структуру молекулы [9]. Поскольку точность предсказательной модели зависит от информативности дескрипторов, их отбор является критически важным этапом.

На сегодняшний день не существует правила выбора оптимальных дескрипторов для определенной задачи. Более распространенный подход заключается в генерации широкого набора дескрипторов и последующей фильтрации полученных данных. Для расчета молекулярных дескрипторов использована библиотека mordred [10] на языке Python, с помощью которой рассчитан 1391 дескриптор. Для устранения проблемы «проклятия размерности» проведена фильтрация с удалением мультиколлинеарных признаков и данных с низкой дисперсией, в результате чего осталось 480 дескрипторов.

Построение модели машинного обучения

Молекулярные дескрипторы и целевые значения свойств подаются в модель МО в качестве входных и выходных значений. Для прогнозирования свойств полимеров использован алгоритм случайного леса (*random forest*) [11], зарекомендовавший себя высокой точностью в подобных задачах [12–14].

Оптимизация гиперпараметров проводилась методом grid search с пятикратной кросс-валидацией. В качестве метрики использовался коэффициент детерминации. Обучение модели проводилось на 80 % данных, тестирование – на оставшихся 20 %. На рисунке 1 представлены значения коэффициента детерминации для целевых свойств.

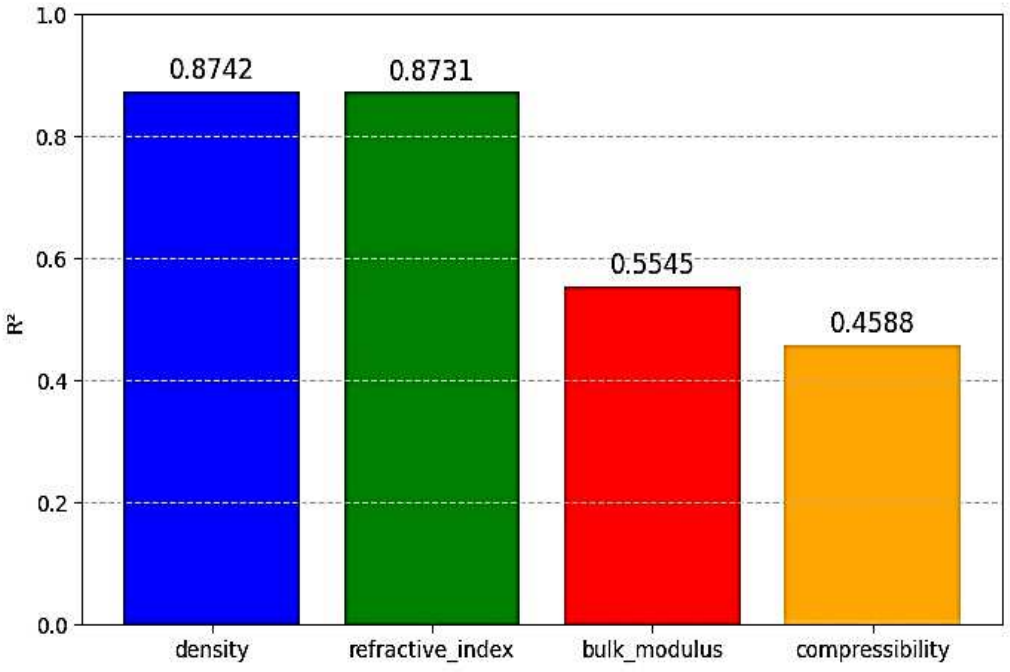

Рисунок 1 – Значение коэффициента детерминации физических свойств полимеров

На рисунке 2 представлен график рассеяния спрогнозированных значений относительно фактических. Чем ближе точки к диагональной линии, тем выше точность модели. Плотность (R^2 = 0.8742) и показатель преломления (R^2 = 0.8731) предсказываются с высокой точностью. Модуль упругости (R^2 = 0.5545) и сжимаемость (R^2 = 0.4588) предсказываются хуже, что может быть связано с отсутствием важных признаков в данных или сложными нелинейными зависимостями.

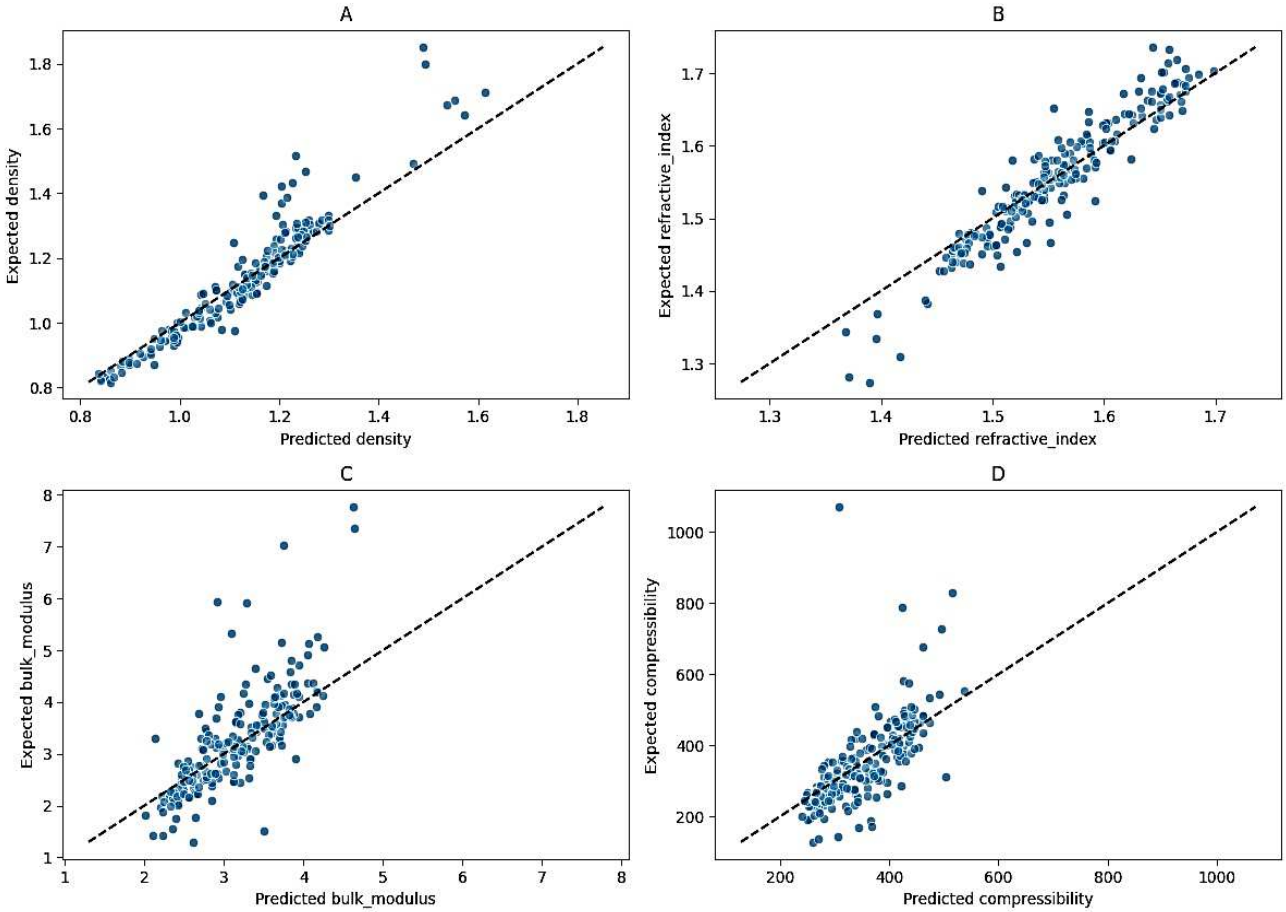

Рисунок 2 – Спрогнозированные (ось X) и фактические (ось Y) значения:
A) плотности, B) показателя преломления, C) модуля упругости, D) сжимаемости

Вывод

Применение молекулярных дескрипторов и машинного обучения показывают высокий потенциал для прогнозирования свойств полимеров. Однако успешность прогнозирующих моделей во многом определяется качеством исходных данных. В частности, для точного прогнозирования механических свойств необходимо учитывать не только параметры, вычисленные с помощью структуры мономера, но и структурные характеристики полимера: степень полимеризации, межмолекулярные взаимодействия, полярность и другие факторы. Несмотря на это, уже на данном этапе модель случайного леса позволяет с высокой точностью предсказывать плотность и показатель преломления с коэффициентом детерминации равным 0.8742 и 0.8731 соответственно, что подтверждает эффективность данного подхода.

Список литературы / List of references:

1. A review on the application of molecular descriptors and machine learning in polymer design / Y. Zhao [et al.] // Polym. Chem. – 2023. – Vol. 14. – № 29. – P. 3325–3346.

2. Polymer design using genetic algorithm and machine learning / C. Kim [et al.] // Comput. Mater. Sci. – 2021. – Vol. 186. – P. 110067.

3. Polymer informatics: Current status and critical next steps / L. Chen [et al.] // Mater. Sci. Eng. R Rep. – 2021. – Vol. 144. 100595.

4. AI Applications through the Whole Life Cycle of Material Discovery / J. Li [et al.] // Matter. – 2020. – Vol. 3. – № 2. – P. 393–432.

5. Accelerating materials discovery using artificial intelligence, high performance computing and robotics / E.O. Pyzer-Knapp [et al.] // Npj Comput. Mater. – 2022. – Vol. 8. – № 1. – P. 84.

6. Dielectric Polymer Property Prediction Using Recurrent Neural Networks with Optimizations / A.L. Nazarova [et al.] // J. Chem. Inf. Model. – 2021. – Vol. 61. – № 5. – P. 2175–2186.

7. Machine-Learning-Assisted De Novo Design of Organic Molecules and Polymers: Opportunities and Challenges / G. Chen [et al.] // Polymers. – 2020. – Vol. 12. – № 1. – P. 163.

8. RadonPy: automated physical property calculation using all-atom classical molecular dynamics simulations for polymer informatics / Y. Hayashi [et al.] // Npj Comput. Mater. – 2022. – Vol. 8. – № 1. – P. 222.

9. Danishuddin Khan A.U. Descriptors and their selection methods in QSAR analysis: paradigm for drug design // Drug Discov. Today. – 2016. – Vol. 21. – № 8. – P. 1291–1302.

10. Mordred: a molecular descriptor calculator / H. Moriwaki [et al.] // J. Cheminformatics. – 2018. – Vol. 10. – № 1. – P. 4.

11. Breiman L. Random Forests // Mach. Learn. – 2001. – Vol. 45. – № 1. – P. 5–32.

12. Flame retardant properties of metal hydroxide-based polymer composites: A machine learning approach / [et al.] Xiao J. // Compos. Commun. – 2023. – Vol. 40. – P.101593.

13. Fatigue life prediction of bending polymer films using random forest / M. Kishino [et al.] // Int. J. Fatigue. – 2023. – Vol. 166. – P. 107230.

14. Gupta P. Dielectric properties of polymer nanocomposite interphases from electrostatic force microscopy using machine learning / P. Gupta, L.S. Schadler, R. Sundararaman // Mater. Charact. – 2021. – Vol. 173. – P. 110909.