

УДК 004.94

Научная статья

<https://doi.org/10.23947/2587-8999-2023-6-1-77-82>**Прогнозирование кинетики сложных процессов люминесценции на Python**С. В. Слепнёв¹, К. Ф. Коледина^{1,2} ✉¹Уфимский государственный нефтяной технический университет, Российская Федерация, г. Уфа, ул. Космонавтов, 1²Институт нефтехимии и катализа УФИЦ РАН, Российская Федерация, г. Уфа, пр. Октября, 141✉ koledinakamila@mail.ru**Аннотация**

Введение. На сегодняшний день полиарилефталиды (ПАФ) находят широкое применение в оптоэлектронике. При этом реакции, протекающие при синтезе полиарилефталидов, имеют сложный характер, который до сих пор не удалось описать с помощью математических моделей. В связи с этим, невозможно использовать ПАФ во многих процессах. При этом ПАФ обладают люминесценцией, хорошими оптическими и электрофизическими свойствами. Выяснение механизмов возникновения люминесцирующих состояний ПАФ представляет как фундаментальный, так и практический интерес. В связи со сложностью расчета кинетики интенсивности свечения полиарилефталидов с помощью известных математических моделей была поставлена цель исследования — построить с помощью методов машинного обучения систему, прогнозирующую значения люминесценции в зависимости от температуры и времени нагревания.

Материалы и методы. Подготовлены к вычислениям экспериментальные данные, обосновано использование методов «случайный лес» и «градиентный бустинг», выбран способ подбора гиперпараметров данных моделей и обоснована целесообразность его использования, построены оптимальные модели и получены предсказания.

Результаты исследования. Разработан алгоритм предсказания интенсивности свечения полиарилефталидов. Используя методы машинного обучения на экспериментальных данных, были определены ключевые гиперпараметры системы и достигнута средняя точность предсказания значений — 80 %.

Обсуждение и заключения. Прогнозы высокой точности позволят предсказывать, как будут реагировать на внешнее воздействие продукты, включающие в свой состав полиарилефталиды. В работе представлено два метода решения задачи, так как они показали наилучшие результаты.

Ключевые слова: Python, люминесценция, методы машинного обучения, случайный лес, градиентный бустинг.

Для цитирования. Слепнёв, С. В. Прогнозирование кинетики сложных процессов люминесценции на Python / С. В. Слепнёв, К. Ф. Коледина // Computational Mathematics and Information Technologies. — 2023. — Т. 6, № 1. — С. 77–82. <https://doi.org/10.23947/2587-8999-2023-6-1-77-82>

Original article

Predicting the kinetics of complex luminescence processes in PythonS. V. Slepnev¹, K. F. Koledina^{1,2}  ✉¹Ufa State Petroleum Technical University, 1, Kosmonavtov str., Ufa, Russian Federation²Institute of Petrochemistry and Catalysis of the Russian Academy of Sciences, 141, October Ave, Ufa, Russian Federation✉ koledinakamila@mail.ru

Abstract

Introduction. Polyarylene phthalides (PAF) are widely used in optoelectronics today. The reactions occurring during the synthesis of polyarylene phthalides have a complex character, which has not yet been described using mathematical models. In this regard, it is impossible to use PAF in many processes. Polyarylene phthalides have luminescence, good optical and electrophysical properties. The elucidation of the mechanisms of the occurrence of luminescent states of PAF is of both fundamental and practical interest.

The elucidation of the mechanisms of the occurrence of luminescent states of PAF is of both fundamental and practical interest. Due to the complexity of calculating the kinetics of the luminescence intensity of polyarylene phthalides using known mathematical models, the aim of the study was to build a system using machine learning methods that predicts luminescence values depending on temperature and heating time.

Materials and methods. Experimental data have been prepared for calculations, the use of “random forest” and “gradient boosting” methods has been justified, a method for selecting hyperparameters of these models has been selected and the expediency of its use has been justified, optimal models have been constructed and predictions have been obtained.

The results of the study. An algorithm for predicting the luminescence intensity of polyarylene phthalides has been developed. Using machine learning methods based on experimental data, the key hyperparameters of the system were determined and the average accuracy of predicting values was achieved — 80 %.

Discussion and conclusions. High-accuracy forecasts will allow predicting how products containing polyarylene phthalides will react to external influences. The paper presents two methods for solving the problem, as they showed the best results.

Keywords: Python, luminescence, machine learning methods, random forest, gradient boosting.

For citation. Slepnev, S. V. Forecasting the kinetics of complex luminescence processes in Python / S. V. Slepnev, K. F. Kaledina // Computational Mathematics and Information Technologies. — 2023. — Vol. 6, no. 1. — P. 77–82. <https://doi.org/10.23947/2587-8999-2023-6-1-77-82>

Введение. На сегодняшний день органические полимерные материалы находят широкое применение в оптоэлектронике. Одной из разновидностей полимеров, пригодных для этих целей, могут быть полиарилефталиды (ПАФ). ПАФ отличаются высокой термо- и хемостойкостью, высокими пленкообразующими свойствами. При этом ПАФ обладают люминесценцией, хорошими оптическими и электрофизическими свойствами. Выяснение механизмов возникновения люминесцирующих состояний ПАФ представляет как фундаментальный, так и практический интерес. Предполагается, что люминесценция ПАФ обусловлена образованием активных интермедиатов при энергетическом воздействии на полимер, однако их химическая природа и свойства совершенно не изучены [1].

Так как реакции, протекающие при синтезе полиарилефталидов, имеют сложный характер, совершенно очевидна необходимость подробных исследований в этой практически неизученной области. Однако, несмотря на исследования и проводимые эксперименты [1–3], так и не была построена математическая модель, описывающая поведение свечения полиарилефталидов.

В связи со сложностью расчета кинетики интенсивности свечения полиарилефталидов с помощью известных математических моделей [2, 3], была поставлена цель — построить с помощью методов машинного обучения систему, прогнозирующую значения люминесценции в зависимости от температуры и времени нагревания.

Для реализации цели были поставлены следующие задачи:

- подготовить экспериментальные данные для расчетов;
- проанализировать алгоритмы машинного обучения и создать программу с применением наиболее оптимальных методов;
- подобрать наиболее удачные гиперпараметры для моделей.

Материалы и методы. Полиарилефталиды — тип ароматических полимеров. Предполагается, что люминесценция полиарилефталидов обусловлена образованием активных интермедиатов при энергетическом воздействии на полимер. В экспериментальных исследованиях температуру нагревания пленки

полиарилфталидов авторы меняли в диапазоне от 298 до 460 К в течение нескольких часов, используя различные скорости контролируемого нагрева и охлаждения пленки полиарилфталидов. Ими было установлено, что температурное воздействие на пленку приводит к возникновению долго затухающего свечения — рекомбинационной люминесценции [1].

Для накопления стабильных ион-радикалов пленку ПАФ облучали нефiltroванным светом в течение 10 минут (лампа 100 Вт), после чего оставляли при температуре 298 К на 8 часов. Температуру пленки ПАФ меняли в диапазоне от 298 до 460 К в течение нескольких часов, используя различные скорости контролируемого нагрева и охлаждения пленки ПАФ.

Данные по характеристикам индивидуальных полиарилфталидов представлены следующими категориями: время, температура, интенсивность свечения. Всего было 20 экспериментов, 200 значений по каждой характеристике.

На полученных данных были опробованы различные методы машинного обучения и выбраны наиболее подходящие для последующего улучшения модели. Подбор гиперпараметров реализован через рандомизированный поиск по параметрам, каждый из которых выбирается из распределения по возможным значениям.

Случайный лес. Алгоритм случайного леса (Random Forest) — универсальный алгоритм машинного обучения, суть которого состоит в использовании ансамбля решающих деревьев. Само по себе решающее дерево предоставляет крайне невысокое качество классификации, но из-за большого их количества результат значительно улучшается [4].

По сравнению с другими методами машинного обучения, теоретическая часть алгоритма Random Forest проста. Нет большого объема теории, необходима только формула итогового классификатора $a(x)$:

$$a(x) = \frac{1}{N} \sum_{i=1}^N b_i(x), \quad (1)$$

где N — количество деревьев; i — счетчик для деревьев; b — решающее дерево; $a(x)$ — сгенерированная на основе данных выборка.

Однако, несмотря на универсальность, у данного метода наблюдается ряд существенных недостатков [5]:

- сложность интерпретации;
- случайный лес не умеет экстраполировать;
- алгоритм склонен к переобучению на сильно зашумленных данных;
- для данных, включающих категориальные переменные с различным количеством уровней, случайные леса предвзяты в пользу признаков с большим количеством уровней;
- большой размер получающихся моделей. Требуется $O(N \cdot C)$ памяти для хранения модели, где C — число деревьев.

Градиентный бустинг. Это метод машинного обучения, который создает решающую модель прогнозирования в виде ансамбля слабых моделей прогнозирования, обычно деревьев решений. Он строит модель поэтапно, позволяя оптимизировать произвольную дифференцируемую функцию потерь [6–8].

Пусть L — дифференцируемая функция потерь, а алгоритм $a(x)$ представляет собой композицию базовых алгоритмов:

$$a(x) = a_k(x) = b_1(x) + \dots + b_k(x), \quad (2)$$

где базовый алгоритм b_k обучается так, чтобы улучшить предсказания текущей композиции:

$$b_k = \arg \min \sum_{i=1}^N L(y_i, a_{k-1}(x_i) + b(x_i)). \quad (3)$$

Модель b_0 выбирается так, чтобы минимизировать потери на обучающей выборке:

$$b_0 = \arg \min \sum_{i=1}^N L(y_i, b(x_i)). \quad (4)$$

Для построения базовых алгоритмов на следующих шагах рассмотрим разложение Тейлора функции потерь L до первого члена в окрестности точки $(y_i, a_{k-1}(x_i))$:

$$L(y_i, a_{k-1}(x_i)) = L(y_i, a_{k-1}(x_i)) + b(x_i) g_i^{k-1}. \quad (5)$$

Избавившись от постоянных членов, можно получить оптимизационную задачу:

$$b_k \approx \arg \min \sum_{i=1}^N b(x_i) g_i^{k-1}. \quad (6)$$

На каждой итерации базовые алгоритмы b_k обучаются предсказывать значения антиградиента функции потерь по текущим предсказаниям композиции.

К основным недостаткам данного метода можно отнести подверженность переобучению и систему голосования оценщиков. Голосование оценщиков в градиентном бустинге происходит неравноправно. Одни оценщики имеют более высокий вес, чем другие. Как правило, голос самого первого обученного оценщика имеет самый минимальный вес, а последний оценщик имеет самый большой вес при голосовании.

Результаты исследования. Решение задачи регрессии осуществлялось с помощью языка программирования Python, а оценка точности решения — через коэффициент детерминации, который рассчитывается по формуле:

$$R^2 = 1 - \frac{S_{res}}{S_{tot}}, \quad (7)$$

где S_{res} — сумма квадратов остаточных ошибок, а S_{tot} — общая сумма ошибок.

На рис. 1, 2 приведены результаты моделирования кинетики интенсивности свечения ПАФ в зависимости от изменения температуры. Красной кривой показаны экспериментальные данные интенсивности свечения после воздействия температуры (синяя пунктирная кривая по вспомогательной вертикальной оси).

На рис. 1 приведено моделирование интенсивности методом градиентного бустинга (зеленая кривая). На рис. 2 приведено моделирование интенсивности методом случайного леса (зеленая кривая).

Коэффициент детерминации, при использовании градиентного бустинга, равен 83 %, а в случае применения случайного леса — 88 %.

Как и предполагалось, результаты получились немного отличными друг от друга в силу различия методов. При использовании метода градиентного бустинга предсказание оказывается менее точным, и пики прогнозируются в те моменты времени, когда их не было. В случае случайного леса имеются предсказания, которые лежат ближе к экспериментальным данным. Поскольку градиентный бустинг справляется с предсказанием температуры хуже, это влияет на предсказательную способность всей модели. Такой результат может быть обоснован «неидеальным» подбором гиперпараметров и скачкообразным изменением интенсивности свечения, что затрудняет определение направления градиента функции.

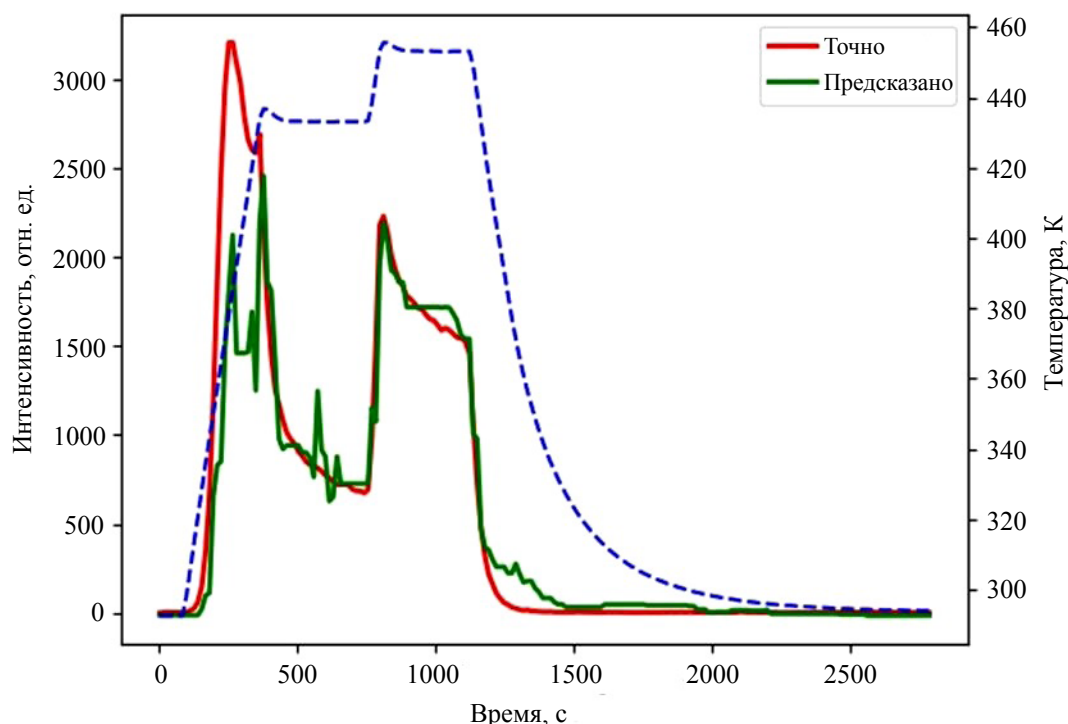


Рис. 1. Моделирование кинетики интенсивности свечения ПАФ методом «Градиентный бустинг»

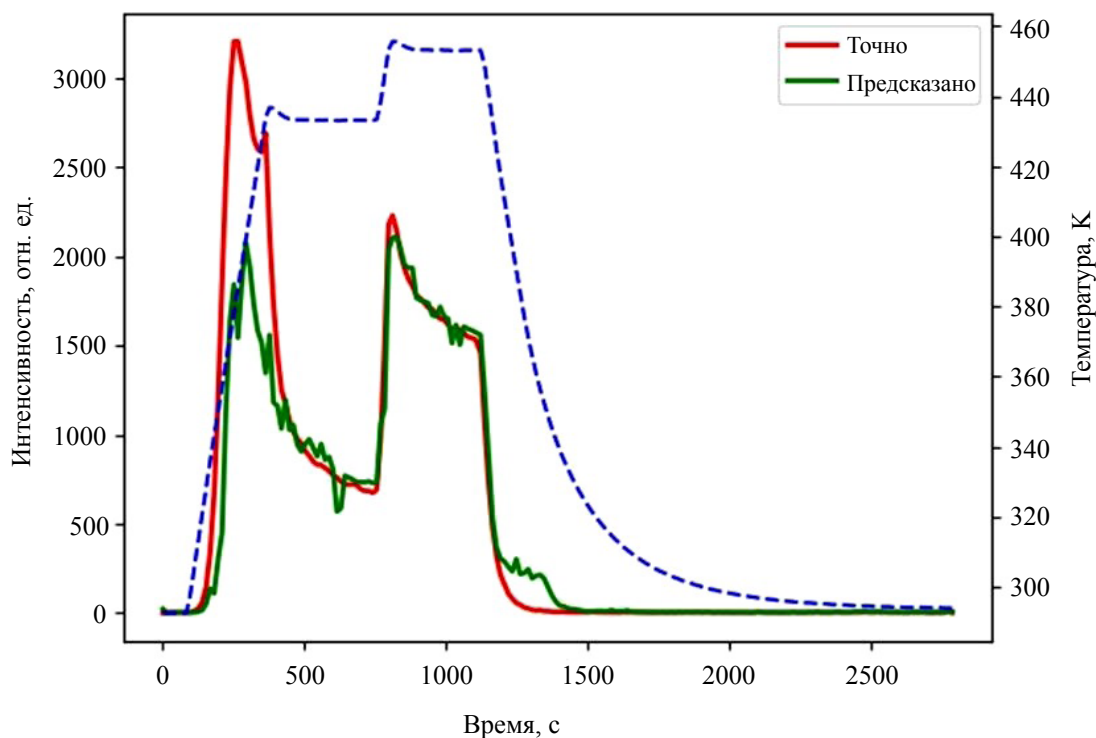


Рис. 2. Моделирование кинетики интенсивности свечения ПАФ методом «Случайный лес»

Обсуждение и заключения. Результат, полученный в ходе решения задачи прогнозирования *интенсивности свечения* полиарилефталидов актуален для применения в промышленных и лабораторных процессах. Несмотря на высокую предсказательную способность данной модели, она имеет свойство часто ошибаться при предсказании пиков, что требует дальнейшей доработки. Для уменьшения погрешности моделирования можно прибегнуть к комбинированию подходов и использованию иных методов для получения большей точности. Дополнительным способом улучшения модели также является более точный подбор гиперпараметров. Он, в свою очередь, требует высоких вычислительных мощностей, поскольку использует перебор всех возможных сочетаний и для текущей задачи может осуществляться неделями. Потенциальным решением проблемы может стать предварительный анализ данных с помощью генетических алгоритмов для поиска экстремумов и последующая передача данных для обучения алгоритмам машинного обучения.

Список литературы

1. Овчинникова, М. Ю. Кинетическая модель термостимулированной люминесценции пленок полидифениленфталидов / М. Ю. Овчинникова, В. А. Антипин, С. Л. Хурсан // Кинетика и катализ. — 2019. — Т. 60, № 5. — С. 547–554.
2. Koledina, K. F. Solving the problem of multi-criteria optimization of the synthesis reaction of benzylalkyl esters by the method of “ideal” point and lexicographic ordering / K. F. Koledina, A. A. Alexandrova // Computational mathematics and information technologies. — 2022. — Vol. 1., no. 1. — P. 12–19. [Doi 10.23947/2587-8999-2022-1-1-12-19](https://doi.org/10.23947/2587-8999-2022-1-1-12-19)
3. Shaimardanova, G. F. Genetic algorithm for solving the inverse problem of chemical kinetics / G. F. Shaimardanova, K. F. Koledina // Computational mathematics and information technologies. — 2022. — Vol. 1, no. 1. — P. 41–49. [Doi 10.23947/2587-8999-2022-1-1-41-49](https://doi.org/10.23947/2587-8999-2022-1-1-41-49)
4. Chistyakov, S. P. Random forests: an overview // Proceedings of the Karelian Scientific Center of the Russian Academy of Sciences. — 2013. — No. 1. — P. 117–136.
5. Брантон С. Л., Куц Дж. Н. Анализ данных в науке и технике / пер. с англ. А. А. Слинкина. — Москва : ДМК Пресс, 2021. — 574 с.
6. Muller, A. Introduction to Machine Learning using Python / A. Muller, S. Guido // O'Reilly. — 2016. — 340 p.
7. Sun, S. An adaptive k-nearest neighbor algorithm / S. Sun, R. Huang // IEEE. — 2010. — Vol. 1. — P. 91–94.

8. Орельен Жерон Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow. — Москва : Диалектика, 2018. — 690 с.

Поступила в редакцию 10.02.2023.

Поступила после рецензирования 02.03.2023

Принята к публикации 03.03.2023.

Об авторах:

Сергей Вячеславович Слепнёв, студент кафедры информационных технологий и прикладной математики Уфимского государственного нефтяного технологического университета (РФ, 450064, г. Уфа, ул. Космонавтов, 1). Кафедра «Информационных технологий и прикладной математики», [ORCID](#), 14b59b59@gmail.com

Коледина Камила Феликсовна, доктор физико-математических наук, доцент кафедры информационных технологий и прикладной математики Уфимского государственного нефтяного технологического университета (РФ, 450064, г. Уфа, ул. Космонавтов, 1), научный сотрудник Института нефтехимии и катализа Российской академии наук (РФ, 450075, г. Уфа, проспект Октября, 141), [ORCID](#), koledinakamila@mail.ru

Заявленный вклад соавторов:

Слепнёв Сергей Вячеславович — разработка программы и реализация алгоритма. Коледина Камила Феликсовна — формализация задачи, тестирование, разработка спецификации.

Конфликт интересов.

Авторы заявляют об отсутствии конфликта интересов.

Все авторы прочитали и одобрили окончательный вариант рукописи.