

**РАЗРАБОТКА И ПАРАЛЛЕЛЬНАЯ ГЕНЕРАЦИЯ
ДИВЕРСИФИЦИРОВАННОГО СИНТЕТИЧЕСКОГО ДАТАСЕТА
КИНЕТИЧЕСКИХ КРИВЫХ ДЛЯ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ
В ХИМИЧЕСКОЙ КИНЕТИКЕ**

© С. В. Слепнев^{1*}, К. Ф. Коледина^{1,2}, С. Н. Коледин³

¹Институт нефтехимии и катализа УФИЦ РАН
Россия, Республика Башкортостан, 450075 г. Уфа, пр. Октября, 141.

²Уфимский государственный нефтяной технический университет
Россия, Республика Башкортостан, 450064 г. Уфа, ул. Космонавтов, 1.

³Финансовый университет при Правительстве Российской Федерации
Россия, 125993 г. Москва, Ленинградский пр., 49.

*Email: 14b59b59@gmail.com

Цель работы – разработка программного генератора синтетических кинетических кривых для обучения и тестирования алгоритмов машинного обучения при структурной идентификации механизмов химических реакций. Подход основан на стохастическом варьировании кинетических параметров, начальных условий, температурных характеристик и уровня шума с последующим численным интегрированием систем дифференциальных уравнений. Реализованы типовые механизмы, включая реакции первого порядка, параллельные, последовательные, обратимые, бимолекулярные и автокаталитические превращения. Для ускорения формирования больших выборок предложена параллельная архитектура на основе Python multiprocessing. Показано, что генератор обеспечивает воспроизводимое получение размеченных чистых и зашумленных траекторий с метаданной о механизме и параметрах модели.

Ключевые слова: химическая кинетика, кинетические кривые, синтетический датасет, машинное обучение, численное интегрирование, multiprocessing, LSODA, структурная идентификация.

Введение

Структурная идентификация механизмов химических реакций по экспериментальным кинетическим кривым остается одной из наиболее трудоемких задач химической кинетики. Для восстановления механизма необходимо не только оценить параметры модели, но и определить саму структуру превращений, число промежуточных веществ, наличие параллельных каналов, обратимости, автокаталитических эффектов и конкурирующих стадий. Применение методов машинного обучения позволяет автоматизировать часть этой процедуры, однако эффективность таких методов существенно ограничивается недостатком размеченных экспериментальных данных [1].

Синтетические датасеты позволяют частично устранить это ограничение, поскольку для каждой траектории заранее известны механизм, параметры, начальные условия, температурные характеристики и уровень шума. В отличие от экспериментальных массивов, синтетические данные можно целенаправленно балансировать по типам механизмов и сложности кривых. При этом ключевой задачей становится не только генерация большого числа траекторий, но и обеспечение их разнообразия, информативности и кинетической осмысленности.

Цель настоящей работы – разработать диверсифицированный генератор синтетических кинетических кривых и параллельную вычислительную архитектуру, позволяющую получать большие размеченные датасеты за приемлемое время. Дополнительно разработан графический просмотрщик, предназначенный для контроля качества траекторий, анализа формы кривых и подготовки иллюстративного материала.

Научная новизна работы заключается в объединении диверсифицированного набора типовых кинетических механизмов, стохастической параметризации, контролируемого зашумления и воспроизводимой параллельной генерации данных в едином программном инструменте, ориентированном на задачи структурной идентификации механизмов реакций.

Методология

Каждая синтетическая траектория описывается системой обыкновенных дифференциальных уравнений, связывающей скорости элементарных или укрупненных стадий с концентрациями реагентов, промежуточных веществ и продуктов. В общем виде модель записывается как [2]

$$\frac{dC}{dt} = f(C, k, T, \theta), \quad (1)$$

где C – вектор концентраций веществ, k – набор кинетических констант, T – температура, θ – дополнительные параметры, определяющие начальные условия, шум и характерные временные масштабы. Численное интегрирование выполнялось с использованием функции `solve_ivp` библиотеки `SciPy` с методом `LSODA`, позволяющим автоматически переключаться между режимами для нежестких и жестких систем.

В генератор включены восемь характерных типов механизмов, отражающих распространенные формы кинетических кривых. Перечень механизмов приведен в *табл. 1*. Такой набор обеспечивает представление как монотонных кривых расхода и накопления, так и траекторий с максимумами промежуточных веществ, плато, сигмоидальными участками и конкурирующими каналами [3].

Таблица 1

Типы кинетических механизмов, реализованные в генераторе

Обозначение	Схема	Кинетический смысл
M1	$A \rightarrow R$	Необратимая реакция первого порядка
M2	$A \rightarrow R; A \rightarrow S$	Параллельные конкурирующие каналы
M3	$A \rightarrow R \rightarrow S$	Последовательная реакция с максимумом промежуточного вещества
M4	$A \leftrightarrow R$	Обратимая реакция с выходом к равновесию
M5	$A + B \rightarrow R$	Бимолекулярная реакция
M6	$A \rightarrow R \rightarrow S$	Последовательная схема с дезактивацией
M7	$A \rightarrow R; R \rightarrow S$	Конкурирующая последовательная схема
M8	$A + R \rightarrow 2R$	Автокаталитический рост

Алгоритм генерации одной траектории

Последовательная генерация одной траектории включает выбор механизма, сэмплирование параметров, формирование начальных концентраций, построение временной сетки, интегрирование системы ОДУ, проверку информативности и наложение шума. На этапе проверки отбрасываются неудачные интегрирования, почти неизменные кривые и траектории, не содержащие характерных признаков выбранного класса механизма [4]. Основные этапы приведены в *табл. 2*.

Таблица 2

Основные этапы генерации синтетической кинетической траектории

Этап	Содержание операции
1	Случайный выбор типа механизма из заданного набора
2	Сэмплирование кинетических параметров и температурных характеристик
3	Генерация начальных концентраций
4	Построение неравномерной временной сетки с повышенной плотностью в начальной области
5	Численное интегрирование системы оду
6	Проверка информативности траектории и отбраковка неудачных случаев
7	Наложение аддитивного гауссовского шума с масштабированием по уровню сигнала
8	Запись траектории, параметров и метаданных

Для повышения разнообразия датасета параметры механизмов выбираются из распределений, обеспечивающих широкий диапазон характерных времен, соотношений скоростей стадий и положений максимумов промежуточных веществ. Лог-равномерное распределение используется для констант, изменяющихся на несколько порядков, а бета-распределения применяются для управляемого задания долей, соотношений ветвления и других ограниченных параметров. Такой подход позволяет генерировать не только типовые гладкие кривые, но и пограничные случаи, важные для обучения устойчивых классификаторов.

Шум задается как аддитивная гауссовская компонента с амплитудой, масштабируемой относительно локального уровня сигнала. Это позволяет моделировать как практически гладкие данные, так и экспериментально реалистичные или сильно искаженные кривые. Чистые и зашумленные траектории сохраняются отдельно, что удобно для оценки устойчивости алгоритмов машинного обучения к экспериментальным ошибкам. Примеры кинетических кривых показаны на *рис. 1* и *рис. 2*.

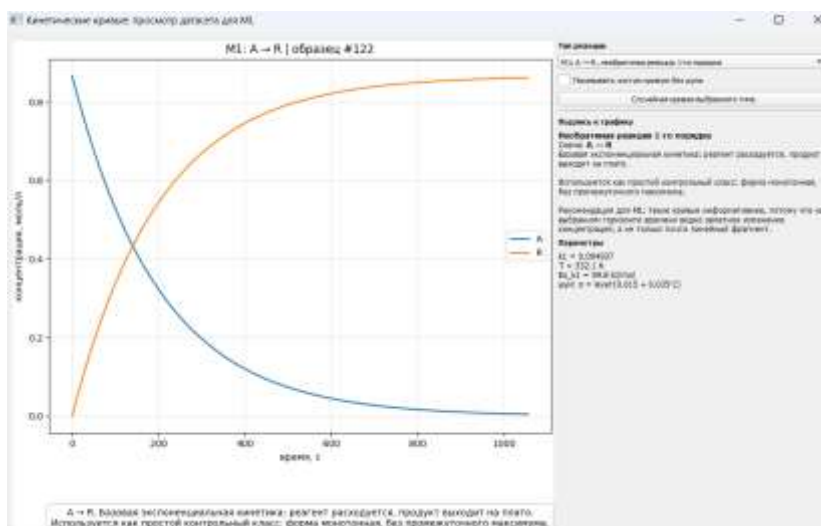


Рис. 1. Пример синтетической кривой для реакции первого порядка $A \rightarrow R$.

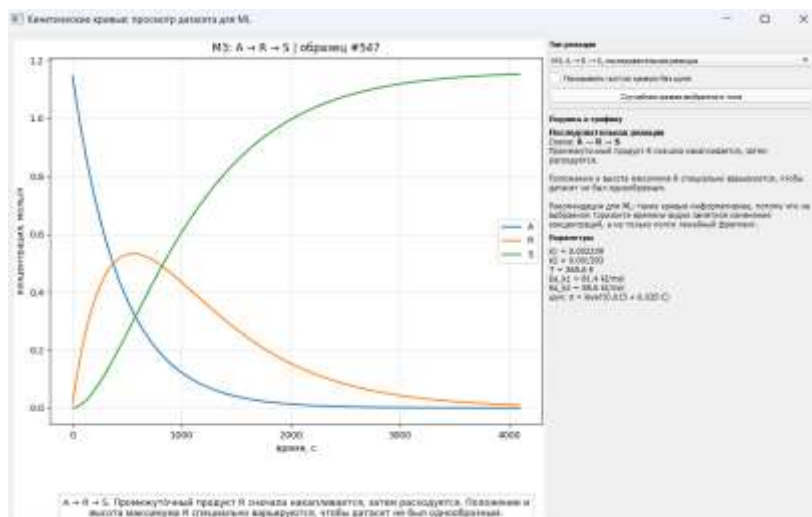


Рис. 2. Пример последовательной схемы $A \rightarrow R \rightarrow S$ с максимумом концентрации промежуточного вещества.

Генерация больших датасетов является вычислительно затратной, поскольку каждая траектория требует численного решения системы ОДУ. Для ускорения расчетов реализована многопроцессная архитектура на основе модуля multiprocessing. Каждая траектория генерируется независимо, поэтому задача хорошо распараллеливается между ядрами процессора. Для хранения и векторной обработки численных массивов используется подход, совместимый с NumPy [5].

Перед запуском рабочих процессов формируется пул уникальных seed-значений. Каждая задача получает собственный seed, что исключает пересечение случайных последовательностей между процессами и обеспечивает воспроизводимость данных при повторном запуске. Использование неупорядоченной выдачи результатов позволяет процессам возвращать готовые траектории сразу после завершения расчета, не ожидая более медленных задач. Передача заданий пакетами снижает накладные расходы межпроцессного взаимодействия.

Поскольку часть траекторий отбраковывается из-за неудачного интегрирования или низкой информативности, число заранее подготовленных seed-значений выбирается с запасом. Это позволяет гарантированно набрать требуемое число качественных образцов даже при высоком проценте отсева. Пример консольного контроля процесса генерации представлен на рис. 3.

```
Generating diverse kinetics: 100% [██████████] | 100000/100000 [03:31<00:00, 471.91it/s]
Saved 100000 trajectories to kinetic_dataset_diverse.h5
Время выполнения: 219.77 секунд
```

Рис. 3. Пример консольного контроля процесса генерации 100 000 кинетических кривых.

Для визуального контроля датасета разработан GUI-просмотрщик, позволяющий выбирать тип реакции, просматривать случайные траектории заданного класса, отображать чистые и зашумленные кривые, а также видеть параметры модели и краткое описание механизма. Такой инструмент облегчает проверку корректности датасета и позволяет быстро выявлять неинформативные или некорректные траектории. Визуализация кинетических кривых выполняется средствами Matplotlib, а табличная организация результатов совместима с pandas [6–7]. Примеры работы просмотрщика приведены на *рис. 4* и *рис. 5*.

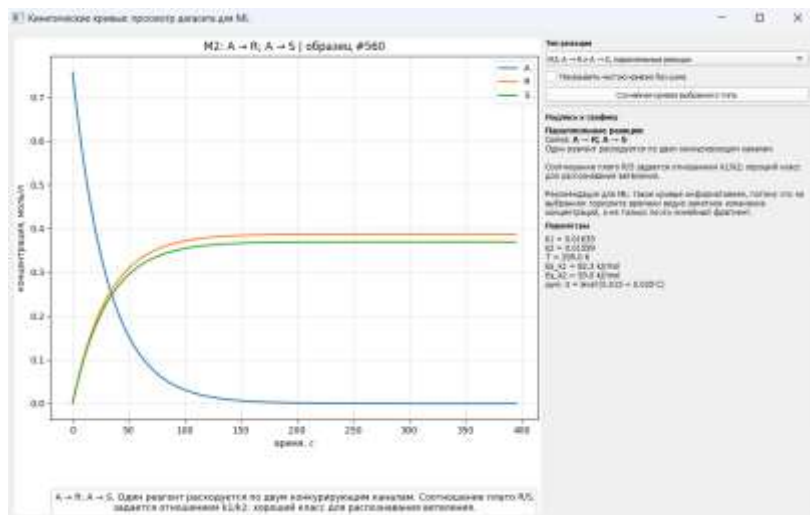


Рис. 4. Пример параллельной реакции $A \rightarrow R$; $A \rightarrow S$ в GUI-просмотрщике.

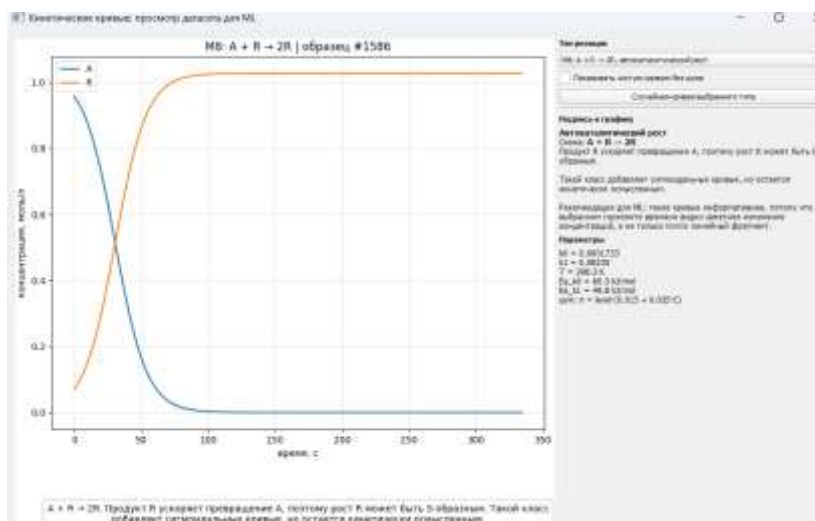


Рис. 5. Пример автокаталитической реакции $A + R \rightarrow 2R$ с S-образной кинетической кривой.

Результаты и обсуждение

Разработанный генератор обеспечивает получение разнообразных кинетических траекторий для восьми классов механизмов. Визуальный анализ кривых показывает, что сформированные данные охватывают монотонные, выходящие на плато, последовательные и автокаталитические формы. Особенно важным является наличие кривых с выраженными максимумами промежуточных веществ, поскольку именно такие признаки часто являются ключевыми для структурной идентификации механизмов.

Параллельная реализация существенно сокращает время генерации. В исходном варианте программы генерация 100 тыс. кинетических кривых на типичной многоядерной рабочей станции занимала порядка 3,5–4 мин. При увеличении числа ядер ускорение близко к линейному до тех пор, пока основная нагрузка приходится на численное интегрирование, а не на ввод–вывод и сериализацию данных.

Получаемый датасет может применяться для обучения классификаторов типа механизма, тестирования устойчивости моделей к шуму, предварительной настройки архитектур нейронных сетей и формирования демонстрационных материалов. Полная информация о параметрах генерации позволяет использовать данные не только как обучающую выборку, но и как воспроизводимый вычислительный эксперимент [8].

Достоверность результатов обеспечивается проверкой успешности численного интегрирования, отбраковкой неинформативных траекторий, сохранением метаинформации о механизме и параметрах модели, а также воспроизводимостью вычислительного эксперимента за счет фиксированных seed-значений. Практическая значимость разработки состоит в возможности формировать обучающие выборки для предварительной настройки и тестирования алгоритмов машинного обучения до применения их к ограниченным экспериментальным данным.

Заключение

В работе разработан диверсифицированный генератор синтетических кинетических кривых для задач машинного обучения в химической кинетике. Реализован набор из восьми характерных механизмов, охватывающий различные формы кинетических траекторий, включая параллельные, последовательные, обратимые и автокаталитические схемы. Предложенная стохастическая параметризация обеспечивает широкий диапазон форм кривых и позволяет управлять уровнем шума.

Разработанная параллельная архитектура на основе multiprocessing обеспечивает эффективную генерацию больших массивов данных с использованием многоядерных процессоров. Применение независимых seed-значений для каждой задачи сохраняет воспроизводимость результатов, а пакетная передача заданий уменьшает накладные расходы. Разработанный GUI-просмотрщик расширяет возможности контроля качества и интерпретации данных.

Полученные результаты могут быть использованы для разработки и тестирования алгоритмов структурной идентификации механизмов химических реакций, в том числе процессов нефтепереработки [9–10], обучения моделей машинного обучения и подготовки иллюстративных материалов для научных публикаций.

Работа выполнена в рамках темы государственного задания Института нефтехимии и катализа УФИЦ РАН «Компьютерное моделирование химических и химико-технологических процессов: разработка и тестирование новых алгоритмов и программ» (FMRS-2025-0031).

ЛИТЕРАТУРА

1. Левеншпиль О. Инженерная кинетика химических реакций. М.: Мир, 1975. [Levenspiel O. Chemical Reaction Engineering. Moscow: Mir, 1975].
2. Fogler H. S. Elements of Chemical Reaction Engineering. 5th ed. Boston: Pearson, 2016.
3. Baerns M., Behr A., Brehm A., et al. Basic Principles in Applied Catalysis. Berlin: Springer, 2004.
4. Virtanen P., Gommers R., Oliphant T. E., et al. SciPy 1.0: fundamental algorithms for scientific computing in Python // Nature Methods. 2020. Vol. 17. P. 261–272.
5. van der Walt S., Colbert S. C., Varoquaux G. The NumPy array: a structure for efficient numerical computation // Computing in Science & Engineering. 2011. Vol. 13. No. 2. P. 22–30.
6. Hunter J. D. Matplotlib: A 2D graphics environment // Computing in Science & Engineering. 2007. Vol. 9. No. 3. P. 90–95.
7. McKinney W. Data structures for statistical computing in Python // Proceedings of the 9th Python in Science Conference. 2010. P. 56–61.
8. Pedregosa F., Varoquaux G., Gramfort A., et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2011. Vol. 12. P. 2825–2830.
9. Zainullin R. Z., Koledina K. F., Akhmetov A. F., Gubaidullin I. M. Kinetics of the Catalytic Reforming of Gasoline // Kinetics and Catalysis. 2017. Vol. 58. No. 3. P. 279–289.
10. Zinnatullin V. F., Koledin S. N., Koledina K. F., Gubaydullin I. M. Integrating neural networks with mathematical programming for gasoline blend formulation // Mathematical Models and Computer Simulations. 2025. Vol. 17. No. S1. P. 48–55.

Поступила в редакцию 07.07.2026 г.

DOI: 10.33184/bulletin-bsu-2026.3.8

**DEVELOPMENT AND PARALLEL GENERATION
OF A DIVERSIFIED SYNTHETIC DATASET
OF KINETIC CURVES FOR MACHINE LEARNING
IN CHEMICAL KINETICS**

© S. V. Slepnev^{1*}, K. F. Koledina^{1,2}, S. N. Koledin³

¹*Institute of Petrochemistry and Catalysis Ufa Federal Research Center of the Russian Academy of Sciences
141 Oktyabrya ave., 450075 Ufa, Republic of Bashkortostan, Russia*

²*Ufa State Petroleum Technological University
1 Kosmonavtov st., 450064 Ufa, Republic of Bashkortostan, Russia.*

³*Financial University under the Government of the Russian Federation
49/2 Leningradsky ave., 125167 Moscow, Russia.*

*Email: 14b59b59@gmail.com

The paper presents a software generator of synthetic kinetic curves for training and testing machine learning algorithms in structural identification of chemical reaction mechanisms. The approach is based on stochastic variation of kinetic parameters, initial conditions, temperature characteristics, and noise levels followed by numerical integration of differential equation systems. The implemented mechanisms include first-order, parallel, consecutive, reversible, bimolecular, and autocatalytic reactions. A parallel architecture based on Python multiprocessing is proposed to accelerate large dataset generation. The generator provides reproducible labelled clean and noisy trajectories with metadata on the mechanism and model parameters.

Keywords: chemical kinetics, kinetic curves, synthetic dataset, machine learning, numerical integration, multiprocessing, LSODA, structural identification.

Received 07.07.2026.

Об авторах / About the authors

СЛЕПНЕВ Сергей Вячеславович

Аспирант лаборатории математической химии.
Институт нефтехимии и катализа УФИЦ РАН.
Email: 14b59b59@gmail.com.

SLEPNEV Sergey Vyacheslavovich

Postgraduate student of the Laboratory of Mathematical Chemistry.
Institute of Petrochemistry and Catalysis
of the Ufa Federal Research Center
of the Russian Academy of Sciences.
Email: 14b59b59@gmail.com
ORCID: 0009-0000-1983-0440.

КОЛЕДИНА Камилла Феликсовна

Доктор физико-математических наук, доцент,
профессор кафедры ИТМ.
Уфимский государственный нефтяной
технический университет.
Старший научный сотрудник лаборатории
математической химии.
Институт нефтехимии и катализа УФИЦ РАН.
Email: koledinakamila@mail.ru.
Web of Science ResearcherID: L-4918-2016.

KOLEDINA Kamila Feliksovna

Doctor of Physical and Mathematical Sciences,
Associate Professor of the Department
of Information Technology and Applied Mathematics.
Ufa State Petroleum Technological University.
Senior Researcher of Laboratory of Mathematical Chemistry.
Institute of Petrochemistry and Catalysis
of the Ufa Federal Research Center
of the Russian Academy of Sciences.
Email: koledinakamila@mail.ru.
ORCID: 0000-0001-8555-0543.
Web of Science ResearcherID: L-4918-2016.
Scopus Author ID: 56005881200.

КОЛЕДИН Сергей Николаевич

Кандидат физико-математических наук,
доцент кафедры информационных технологий.
Финансовый университет
при Правительстве Российской Федерации.
Email: koledinsrg@gmail.com.
Web of Science ResearcherID: AAH-2475-2021.

KOLEDIN Sergey Nikolaevich

Candidate of Physical and Mathematical Sciences,
Associate Professor of the Department of Information Technology.
Financial University under the Government
of the Russian Federation.
Email: koledinsrg@gmail.com.
ORCID: 0000-0003-3291-9794.
Web of Science ResearcherID: AAH-2475-2021.
Scopus Author ID: 57193439673.