

## **ФИЗИЧЕСКИ ОБОСНОВАННЫЕ АТОМНЫЕ ДЕСКРИПТОРЫ: ПРЕДСКАЗАНИЕ $\sigma$ -ПРОФИЛЕЙ МОДЕЛИ COSMO МЕТОДАМИ ГЛУБОКОГО ОБУЧЕНИЯ**

**Шестун П.А.<sup>1</sup>, Кокорина М.С.<sup>1</sup>, Карпушкина И. А.<sup>1</sup>**

**Научный руководитель – Губина Н. В.<sup>1</sup>**

<sup>1</sup>Университет ИТМО

pavelshestun@yandex.ru

Работа выполнена в рамках темы НИР №645123 «ASPEN: физически-обоснованные атомные эмбединги».

### **Введение**

Современные вызовы в химии и фармакологии диктуют переход от анализа молекулы как целого к детальному изучению её отдельных атомных центров. Понимание локальной электронной конфигурации становится фундаментом для точного прогнозирования реакционной способности, поиска сайтов метаболизма и обоснования селективности химических процессов [1]. Несмотря на успехи графовых нейронных сетей (GNN), их архитектура зачастую опирается на формальные атрибуты — атомные номера или типы гибридизации. Подобные дескрипторы являются эмпирическими и лишены глубокого физического смысла, что порождает проблему плохой интерпретируемости моделей и требует избыточных объемов обучающих выборок [1]. В настоящем исследовании мы предлагаем альтернативный путь: формирование физически обоснованных атомных представлений через декомпозицию  $\sigma$ -профилей, полученных в рамках квантово-химического формализма COSMO [2]. Это позволяет наделить эмбединги атомов четким физическим содержанием, отражающим реальное распределение электронной плотности.

### **Основная часть**

Геометрический фундамент модели COSMO, где молекулярная поверхность конструируется как совокупность сфер Ван-дер-Ваальса, создает естественные предпосылки для партиционирования интегрального  $\sigma$ -профиля на индивидуальные атомные составляющие. Такая декомпозиция позволяет преобразовать глобальные данные о распределении экранирующего заряда в локализованные дескрипторы, которые выступают в роли физически прозрачных эмбедингов, чувствительных к химическому окружению. Для генерации обучающей выборки был реализован вычислительный протокол автоматизированной обработки файлов, экспортированных из пакетов xTB и ORCA [3]. Ключевые этапы алгоритма включают идентификацию поверхностных сегментов и их атрибуцию к конкретным центрам, вычисление плотности  $\sigma$ -зарядов и процедуру изотропного сглаживания по схеме Mullins averaging. Завершается препроцессинг аппроксимацией данных на дискретную  $\sigma$ -сетку с учетом весовых коэффициентов площадей. Сравнительный анализ эффективности графовых нейронных сетей показал, что стандартные архитектуры (GCN, GAT, GIN) демонстрируют ограниченную прогностическую способность ( $R^2 = 0,54$ ). Качественный прорыв был достигнут за счет внедрения концепции сетей Колмогорова-Арнольда (KAN) [4]. В отличие от классических многослойных перцептронов, KAN заменяет фиксированные функции активации в узлах на обучаемые одномерные сплайны, расположенные на ребрах сети. Это позволило модели более гибко улавливать нелинейные зависимости в электронном распределении, в результате чего точность предсказания атомных  $\sigma$ -профилей достигла значения  $R^2 = 0,85$ .

## Выводы

Резюмируя, предложенная методология подтверждает возможность оперативной и прецизионной аппроксимации атомных дескрипторов, устраняя потребность в ресурсоемких квантово-химических расчетах *ab initio* для каждой новой структуры. Нам удалось продемонстрировать, что сочетание физического формализма COSMO с гибкостью KAN-графовых архитектур позволяет сохранить строгую интерпретируемость модели при сохранении высокой скорости инференса. Разработанный вычислительный протокол представляет существенный интерес для задач высокопроизводительного скрининга в фармацевтике и материаловедении, где критически важна детализация локальных электронных свойств без потери вычислительной эффективности.

## Литература

1. Gilmer J., Schoenholz S. S., Riley P. F., Vinyals O., Dahl G. E. Neural message passing for quantum chemistry // Proceedings of the 34th International Conference on Machine Learning (ICML). 2017. Vol. 70. P. 1263–1272.
2. Klamt A. Conductor-like screening model for real solvents: A new approach to the quantitative calculation of solvation phenomena // The Journal of Physical Chemistry. 1995. Vol. 99, no. 19. P. 2224–2235.
3. Neese F. Software update: the ORCA program system, version 5.0 // Wiley Interdisciplinary Reviews: Computational Molecular Science. 2022. Vol. 12, no. 5. P. e1606.
4. Liu Z., Wang Y., Vaidya S., Ruehle F., Halverson J., Soljačić M., ..., Tegmark M. KAN: Kolmogorov-Arnold Networks [Электронный ресурс] // arXiv preprint. 2024.