Гибридные подходы на основе машинного обучения и графовых нейронных сетей для точного предсказания температур плавления органических кристаллов

Н. С. Дубонос*, А. С. Новиков‡, Г. В. Боков§, А. А. Митрофанов¶

Предсказание температур плавления критически важно для химии и фармацевтики, но существующие методы сталкиваются с проблемами точности и вычислительной эффективности. В работе собран кодифицированный набор данных из 117 752 органических кристаллов. Проведено сравнение моделей машинного обучения на молекулярных эмбеддингах и графовых нейронных сетей. Гибридная модель TabM с конкатенированными эмбеддингами ChemBERTa и MoLFormer достигает MAE 26.75 K за 27 минут обучения, обеспечивая оптимальный баланс точности и вычислительных затрат по сравнению с Chemprop (MAE 26.44 K, 45 минут).

Ключевые слова: температура плавления, органические кристаллы, машинное обучение, графовые нейронные сети, ChemBERTa, MoLFormer, TabM, Chemprop.

1. Введение и постановка задачи

Предсказание температур плавления органических кристаллов является критически важной задачей для дизайна материалов и фармацевтики. Су-

Дубонос Николай Сергеевич — студент факультета механики и математики МГУ, e-mail: nikolai.dubonos@math.msu.ru.

Dubonos Nikolai Sergeevich — Student, Faculty of Mechanics and Mathematics, Lomonosov Moscow State University.

‡*Новиков Александр Сергеевич* — к.х.н., с.н.с. химического факультета МГУ, e-mail: novikov@radio.chem.msu.ru.

Novikov Alexander Sergeevich — Candidate of Chemical Sciences, Senior Researcher, Faculty of Chemistry, Lomonosov Moscow State University.

§*Боков Григорий Владимирович* — к.ф.-м.н., доцент факультета механики и математики МГУ, e-mail: grigoriy.bokov@math.msu.ru.

Bokov Grigoriy Vladimirovich — Candidate of Physical and Mathematical Sciences, Associate Professor, Faculty of Mechanics and Mathematics, Lomonosov Moscow State University.

¶*Митрофанов Артем Александрович* — к.х.н., с.н.с. Института искусственного интеллекта МГУ, e-mail: a.mitrofanov@iai.msu.ru.

Mitrofanov Artem Alexandrovich — Candidate of Chemical Sciences, Senior Researcher, MSU Institute for Artificial Intelligence.

Опубликовано на условиях лицензии CC-BY 4.0
Published under CC-BY 4.0

ществующие вычислительные подходы можно разделить на три группы: классическое машинное обучение (ML) на основе молекулярных дескрипторов; графовые нейронные сети (GNN), моделирующие топологию молекул; трансформерные архитектуры, обрабатывающие SMILES-строки как последовательности.

Основная проблема заключается в том, что температура плавления – это макроскопическое свойство, зависящее от упаковки кристаллической решетки. Связь между структурой отдельной молекулы и температурой плавления носит разрывный и нелинейный характер. Цель данной работы – разработать гибридный фреймворк, который объединит масштабируемость классического ML и выразительную силу GNN и трансформеров для достижения высокой точности при разумных вычислительных затратах.

2. Подготовка набора данных

Для обучения и тестирования моделей был собран кодифицированный набор данных, объединяющий три открытых источника (Bradley, OPERA и компиляцию Song et al. [1]). В ходе проверки были удалены дубликаты, унифицированы единицы измерения, разрешены конфликты измерений и отфильтрованы соединения с разбросом температур плавления более 5 K. Итоговый набор данных содержит **117 752** уникальных органических кристалла. Анализ сходства Танимото с эталонной базой ChEMBL показал широкое разнообразие химического пространства набора данных.

3. Методы и модели

В работе исследовались четыре типа молекулярных представлений: топологические отпечатки Моргана, 3D-геометрические эмбеддинги Uni-Mol, а также векторы, полученные с помощью предобученных трансформеров ChemBERTa [2] и MoLFormer [3].

Полученные эмбеддинги подавались на вход двум классическим архитектурам для табличных данных: градиентному бустингу **XGBoost** и глубокой нейросети **TabM** [4], эффективно имитирующей ансамбль многослойных перцептронов.

Параллельно была реализована собственная **GNN** на базе сверток GATv2 с механизмом внимания и проведено сравнение с моделями уровня SOTA: **Chemprop** [5], **ALIGNN** и **MEGNet**. Также была выполнена процедура дообучения (fine-tuning) модели MoLFormer.

4. Результаты и обсуждение

Оценка проводилась на фиксированной тестовой выборке (10% данных). Результаты лучших моделей представлены на рис. 1.

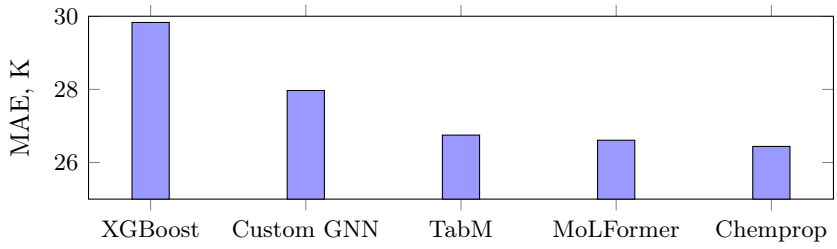


Рис. 1: Сравнение средней абсолютной ошибки (MAE) лучших моделей на тестовой выборке.

Модель XGBoost продемонстрировала высочайшую вычислительную эффективность, но уступила в точности (MAE = 29.83 K). Собственная GNN с механизмом внимания GATv2 улучшила результат до 27.97 K, однако потребовала около 22 часов на обучение.

Наилучшие результаты показали трансформерные и гибридные подходы. Дообученный **MoLFormer** достиг MAE 26.61 K (время обучения 182 мин). Модель **Chemprop** [5] показала самую высокую точность – MAE 26.44 K и RMSE 34.67 K (время обучения 45 мин).

Ключевым результатом работы стало применение гибридного подхода: конкатенация эмбеддингов **ChemBERTa и MoLFormer** с последующей подачей в модель **TabM** [4]. Данная гибридная модель достигла MAE 26.75 K и RMSE 35.36 K. При этом время её обучения составило всего **27 минут**. SHAP-анализ показал, что оба типа эмбеддингов вносят примерно равный вклад в предсказание.

Список литературы

[1] S. Song, Y. Wang, X. Tian, W. Tian, F. Chen, J. Wu, Q. Zhang, “Predicting the Melting Point of Energetic Molecules Using a Learnable Graph Neural Fingerprint Model”, *J. Phys. Chem. A*, **127**:19 (2023), 4328–4337.

[2] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, B. Ramsundar, “ChemBERTa-2: Towards Chemical Foundation Models”, arXiv: 2209.01712.

[3] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, P. Das, “Large-Scale Chemical Language Representations Capture Molecular Structure and Properties”, arXiv: 2106.09553.

[4] Y. Gorishniy, A. Kotelnikov, A. Babenko, “TabM: Advancing Tabular Deep Learning with Parameter-Efficient Ensembling”, *The Thirteenth International Conference on Learning Representations*, 2025.

[5] E. Heid, K. P. Greenman, Y. Chung, S.-C. Li, D. E. Graff, F. H. Vermeire, H. Wu, W. H. Green, C. J. McGill, “Chemprop: A Machine Learning Package for Chemical Property Prediction”, *J. Chem. Inf. Model.*, **64**:1 (2024), 9–17 .

Статья поступила 24 октября 2026 г.

Hybrid Machine Learning and Graph Neural Network Approaches for Accurate Melting Point Prediction of Organic Crystals

N. S. Dubonos, A. S. Novikov, G. V. Bokov, A. A. Mitrofanov

Melting point prediction is crucial for chemistry and pharmaceuticals, but current methods face accuracy and efficiency challenges. We curated a harmonized dataset of 117,752 organic crystals. We compared machine learning models using molecular embeddings and graph neural networks. A hybrid TabM model with concatenated ChemBERTa and MoLFormer embeddings achieves MAE of 26.75 K in 27 minutes of training, offering an optimal balance of accuracy and computational cost compared to Chemprop (MAE 26.44 K, 45 minutes).

Keywords: melting point, organic crystals, machine learning, graph neural networks, ChemBERTa, MoLFormer, TabM, Chemprop.

References

[1] S. Song, Y. Wang, X. Tian, W. Tian, F. Chen, J. Wu, Q. Zhang, “Predicting the Melting Point of Energetic Molecules Using a Learnable Graph Neural Fingerprint Model”, *J. Phys. Chem. A*, **127**:19 (2023), 4328–4337.

[2] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, B. Ramsundar, “ChemBERTa-2: Towards Chemical Foundation Models”, arXiv: 2209.01712.

[3] J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, P. Das, “Large-Scale Chemical Language Representations Capture Molecular Structure and Properties”, arXiv: 2106.09553.

[4] Y. Gorishniy, A. Kotelnikov, A. Babenko, “TabM: Advancing Tabular Deep Learning with Parameter-Efficient Ensembling”, *The Thirteenth International Conference on Learning Representations*, 2025.

[5] E. Heid, K. P. Greenman, Y. Chung, S.-C. Li, D. E. Graff, F. H. Vermeire, H. Wu, W. H. Green, C. J. McGill, “Chemprop: A Machine Learning Package for Chemical Property Prediction”, *J. Chem. Inf. Model.*, **64**:1 (2024), 9–17.

Received on October 24, 2026