Предсказание коэффициентов поглощения в сплавах

Вандышев Георгий (М02-004и)

30 июня 2025 г.

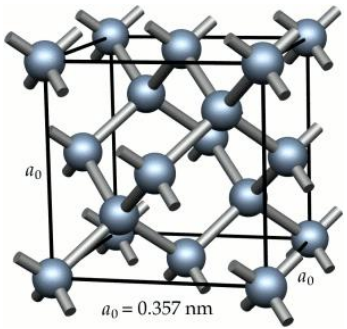
МФТИ

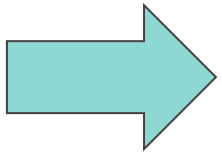
План

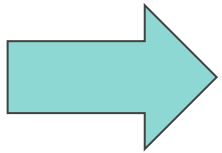
1. Задача
2. База данных
3. Первые шаги
4. Графовые нейросети
5. Еще GNN
6. Возвращение к истокам

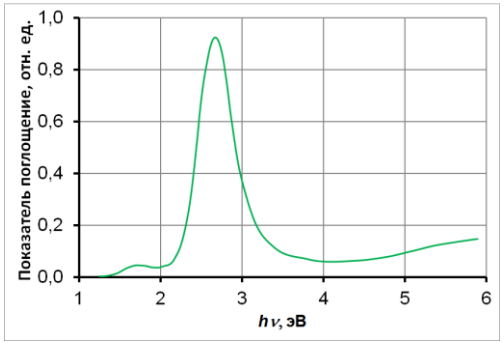
Задача

Построение модели (ML), которая предсказывает по атомной структуре коэффициенты поглощения на длине волны 755 нм, 1064 нм и 1500 нм





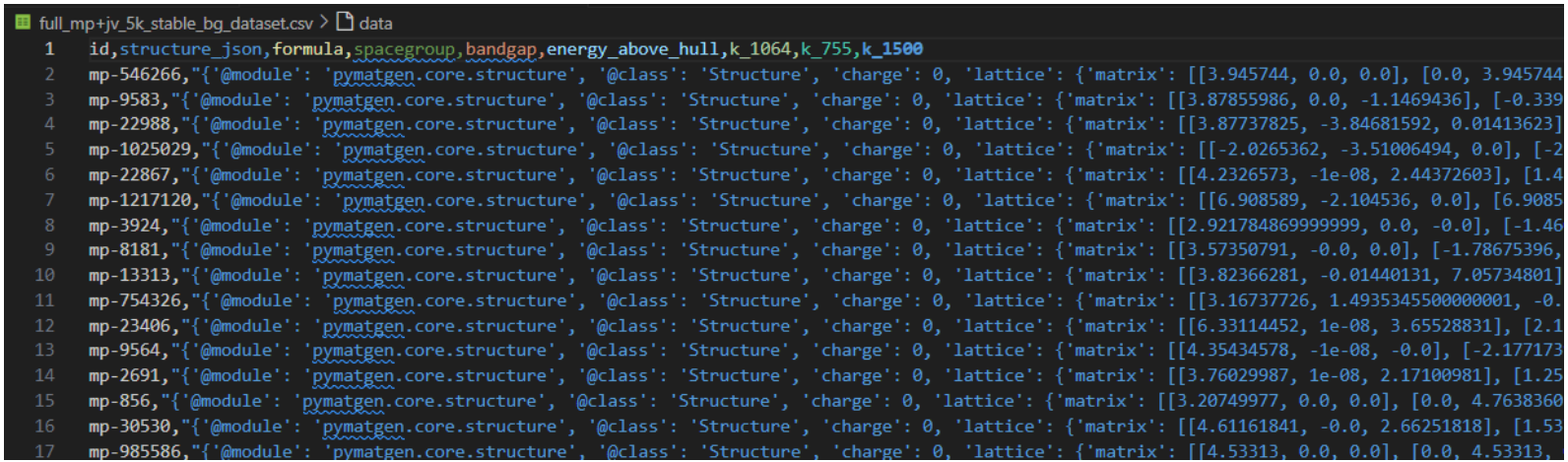




База данных

База данных full_mp+jv_5k_stable_bg_dataset.csv

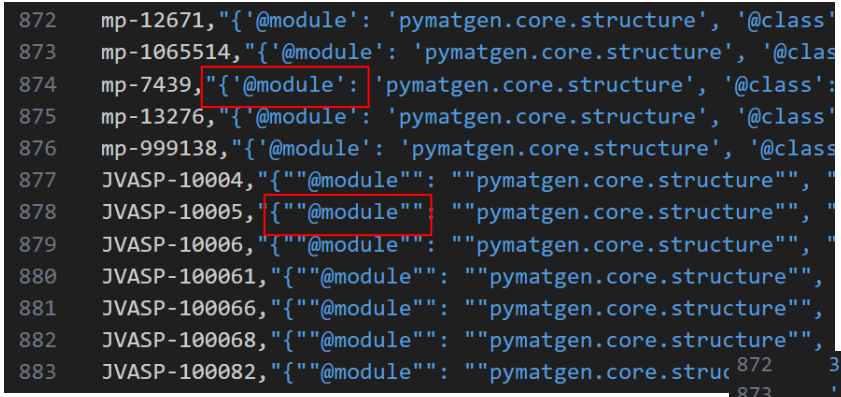
1. По сплавам (видимо из Material Project + JARVIS)
2. Есть показатели поглощения k_755, k_1064, k_1500



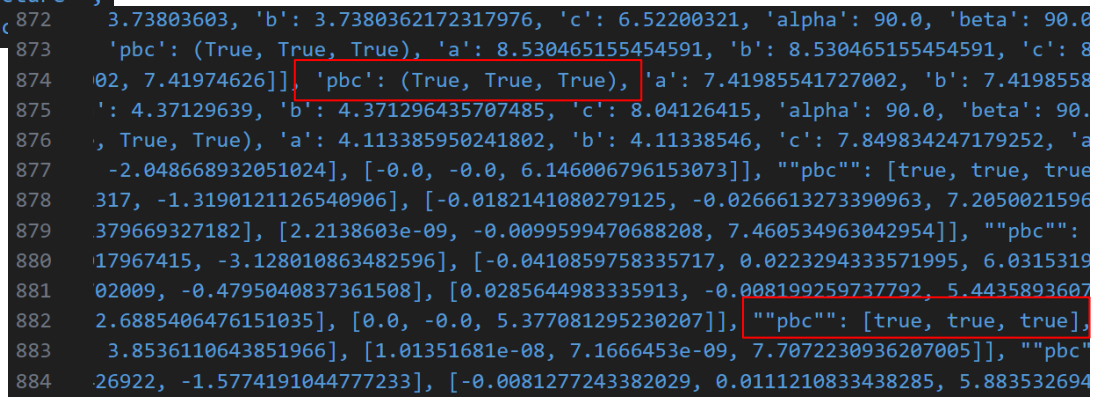
full_mp+jv_5k_stable_bg_dataset.csv > data

```
id,structure_json,formula,spacegroup,bandgap,energy_above_hull,k_1064,k_755,k_1500
mp-546266,"{'@module': 'pymatgen.core.structure', '@class': 'Structure', 'charge': 0, 'lattice': {'matrix': [[3.945744, 0.0, 0.0], [0.0, 3.945744
mp-9583,"{'@module': 'pymatgen.core.structure', '@class': 'Structure', 'charge': 0, 'lattice': {'matrix': [[3.87855986, 0.0, -1.1469436], [-0.339
mp-22988,"{'@module': 'pymatgen.core.structure', '@class': 'Structure', 'charge': 0, 'lattice': {'matrix': [[3.87737825, -3.84681592, 0.01413623]
mp-1025029,"{'@module': 'pymatgen.core.structure', '@class': 'Structure', 'charge': 0, 'lattice': {'matrix': [[-2.0265362, -3.51006494, 0.0], [-2
mp-22867,"{'@module': 'pymatgen.core.structure', '@class': 'Structure', 'charge': 0, 'lattice': {'matrix': [[4.2326573, -1e-08, 2.44372603], [1.4
mp-1217120,"{'@module': 'pymatgen.core.structure', '@class': 'Structure', 'charge': 0, 'lattice': {'matrix': [[6.908589, -2.104536, 0.0], [6.9085
mp-3924,"{'@module': 'pymatgen.core.structure', '@class': 'Structure', 'charge': 0, 'lattice': {'matrix': [[2.921784869999999, 0.0, -0.0], [-1.46
mp-8181,"{'@module': 'pymatgen.core.structure', '@class': 'Structure', 'charge': 0, 'lattice': {'matrix': [[3.57350791, -0.0, 0.0], [-1.78675396,
mp-13313,"{'@module': 'pymatgen.core.structure', '@class': 'Structure', 'charge': 0, 'lattice': {'matrix': [[3.82366281, -0.01440131, 7.05734801]
mp-754326,"{'@module': 'pymatgen.core.structure', '@class': 'Structure', 'charge': 0, 'lattice': {'matrix': [[3.16737726, 1.4935345500000001, -0.
mp-23406,"{'@module': 'pymatgen.core.structure', '@class': 'Structure', 'charge': 0, 'lattice': {'matrix': [[6.33114452, 1e-08, 3.65528831], [2.1
mp-9564,"{'@module': 'pymatgen.core.structure', '@class': 'Structure', 'charge': 0, 'lattice': {'matrix': [[4.35434578, -1e-08, -0.0], [-2.177173
mp-2691,"{'@module': 'pymatgen.core.structure', '@class': 'Structure', 'charge': 0, 'lattice': {'matrix': [[3.76029987, 1e-08, 2.17100981], [1.25
mp-856,"{'@module': 'pymatgen.core.structure', '@class': 'Structure', 'charge': 0, 'lattice': {'matrix': [[3.20749977, 0.0, 0.0], [0.0, 4.7638360
mp-30530,"{'@module': 'pymatgen.core.structure', '@class': 'Structure', 'charge': 0, 'lattice': {'matrix': [[4.61161841, -0.0, 2.66251818], [1.53
mp-985586,"{'@module': 'pymatgen.core.structure', '@class': 'Structure', 'charge': 0, 'lattice': {'matrix': [[4.53313, 0.0, 0.0], [0.0, 4.53313,
```

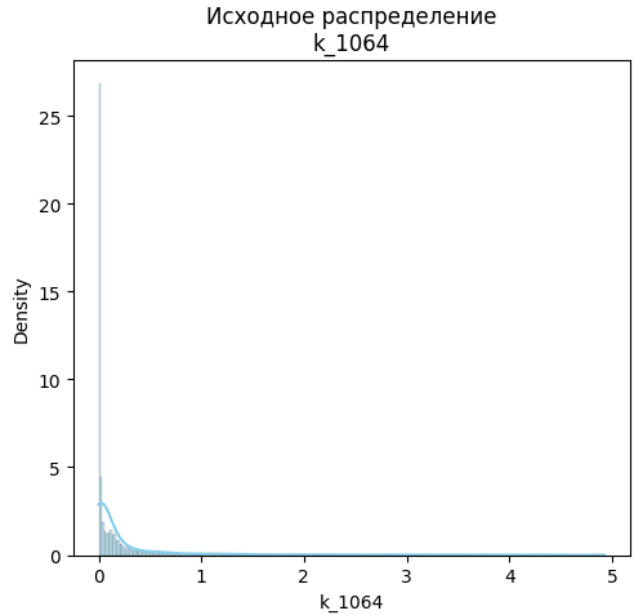
База данных full_mp+jv_5k_stable_bg_dataset.csv

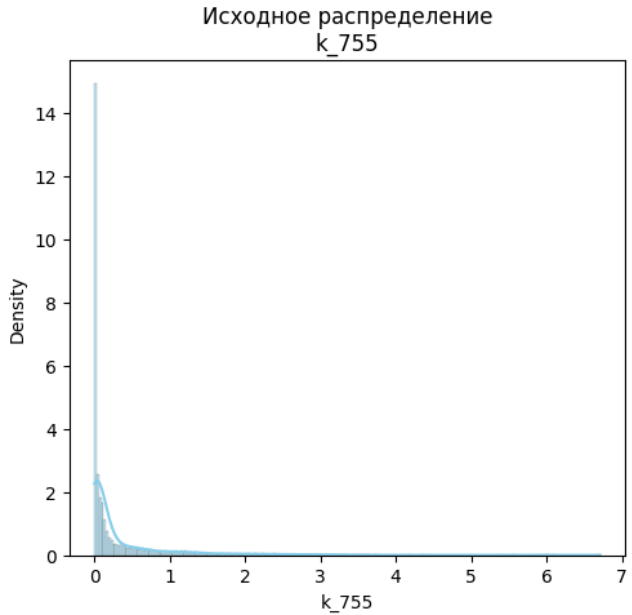


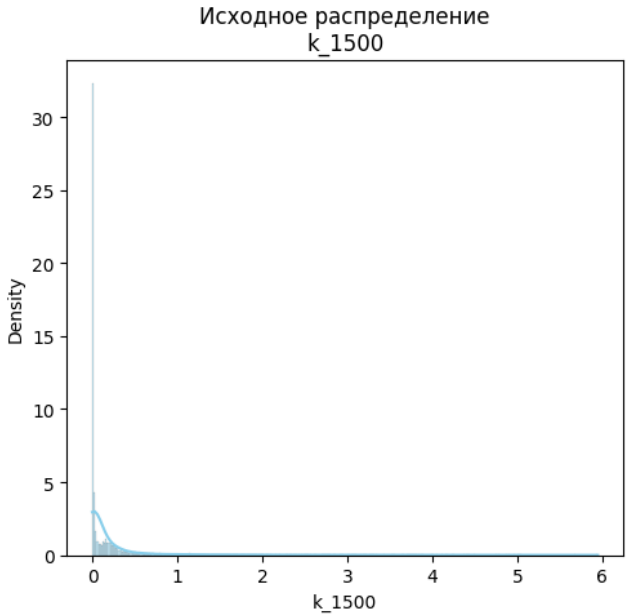
Надо изменить по формату JVASP



Распределение сплавов по k







Первые шаги

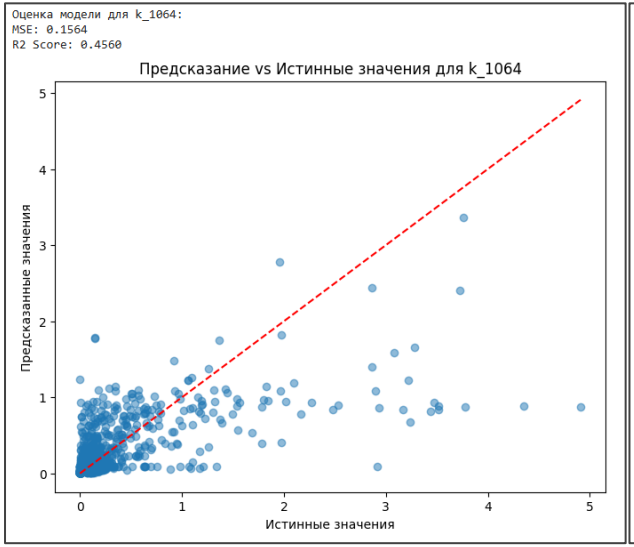
RandomForest.ipynb

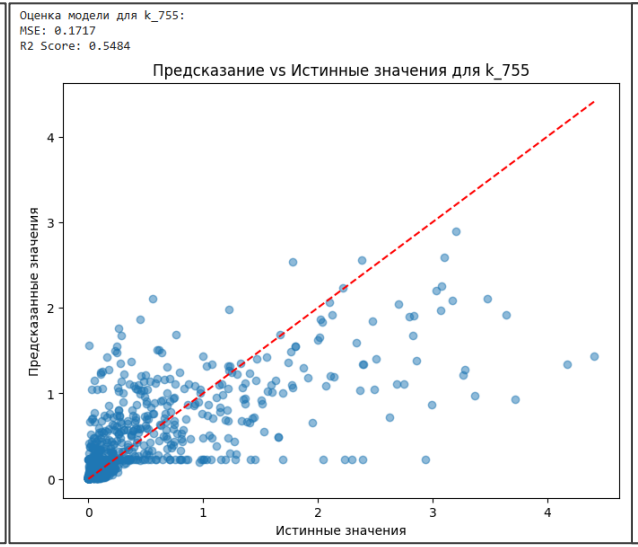
Первые шаги

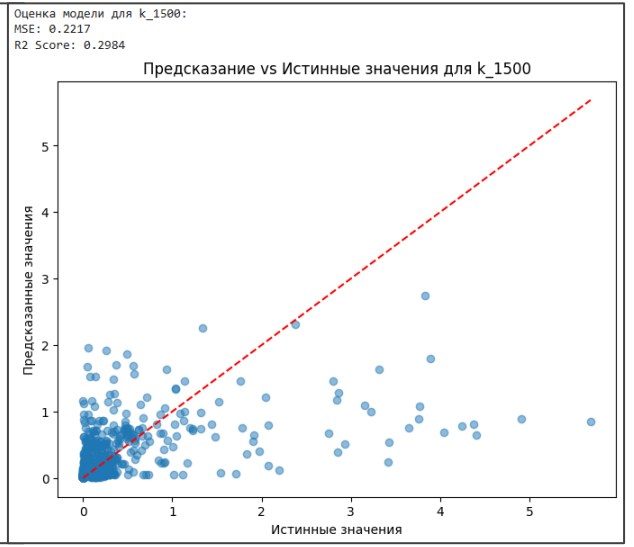
1. В начале было решено протестировать классическую ML модель - RandomForest
2. Из таблицы “full_mp+jv_5k_stable_bg_dataset.csv” были получены свойства сплавов (базовые)
3. Была обучена модель в библиотеке sklearn для каждого значения k
4. Были исследованы гиперпараметры модели -> получены наилучшие

```
feature_names = [
    'volume', 'bandgap', 'energy_above_hull',
    'a', 'b', 'c', 'alpha', 'beta', 'gamma',
    'num_elements', 'spacegroup'
]
```

Результаты для RandomForest







Все не очень хорошо

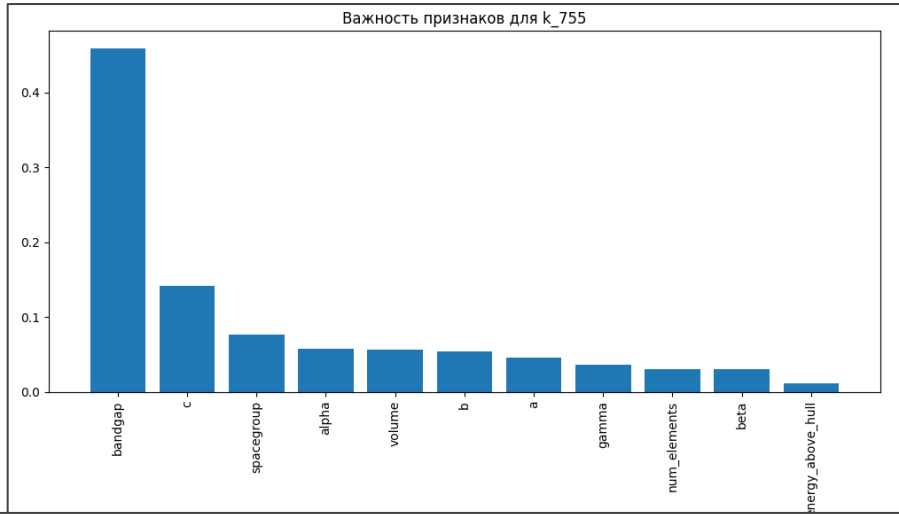
Пример вывода GridSearch:

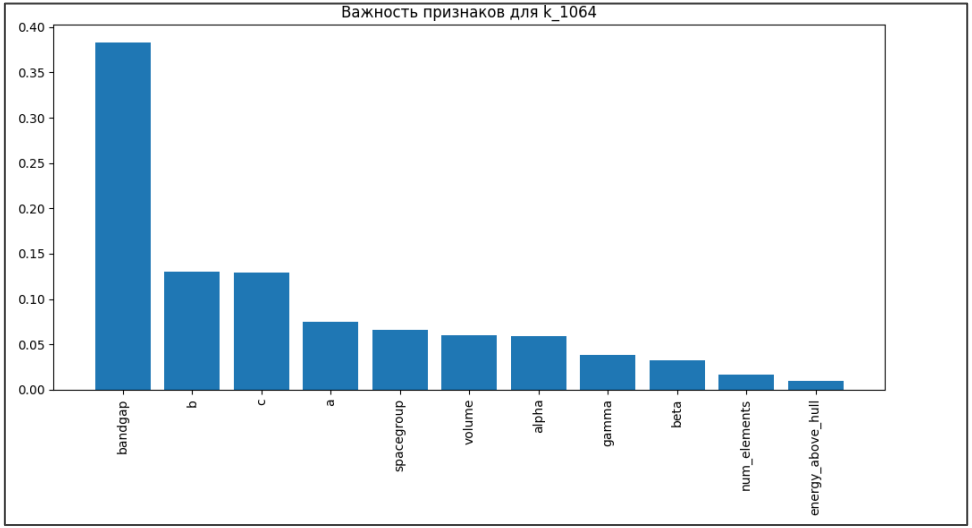
```
{'max_depth': 11, 'max_leaf_nodes': 250, 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 500} -0.19925332709743784
```

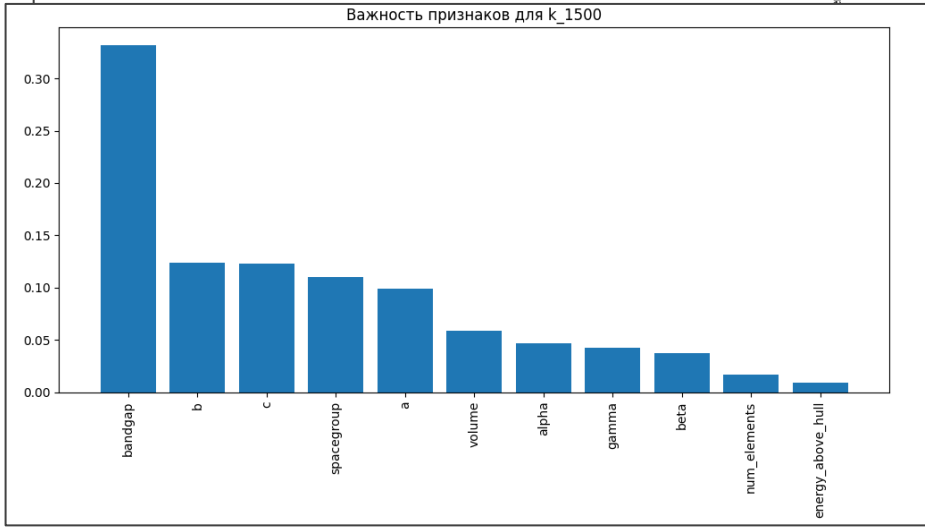
Важность признаков

Самый важный признак - band gap

Далее параметры кристаллической решетки





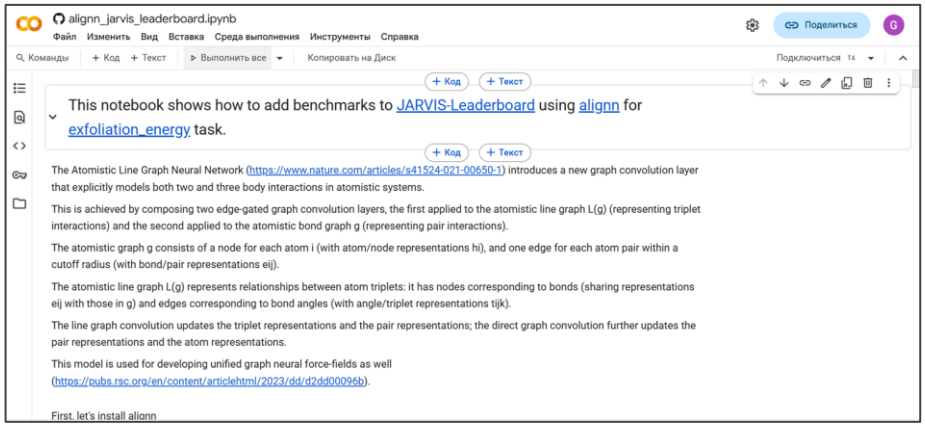


Графовые нейросети

Data_for_AIGNN.ipynb + ALIGNN.ipynb

Графовые нейросети

1. Выбрана ALIGNN
2. Найден ноутбук в качестве примера (официальный - git)
3. Входные параметры = POSCAR файлы VASP
4. Перевод таблицы в папку с входными файлами (alignn/alignn_data)
5. config_k.json - config файл чуть-чуть изменен



https://colab.research.google.com/github/knc6/jarvis-tools-notebooks/blob/master/jarvis-tools-notebooks/alignn_jarvis_leaderboard.ipynb

Проблемы

1. Не полная документация
2. Использует не все конфигурации (через раз)

```
data range 4.92489910736036 4.5216015119897497e-05
line_graph True
100% 6868/6868 [02:41<00:00, 42.40it/s]
data range 3.720222773061368 0.0
line_graph True
100% 858/858 [00:24<00:00, 34.63it/s]
data range 2.9435545702461887 0.0003154817802712
line_graph True
100% 858/858 [00:18<00:00, 45.65it/s]
n_train: 6868
n_val  : 858
n_test : 858
rank 0
world_size 1
```

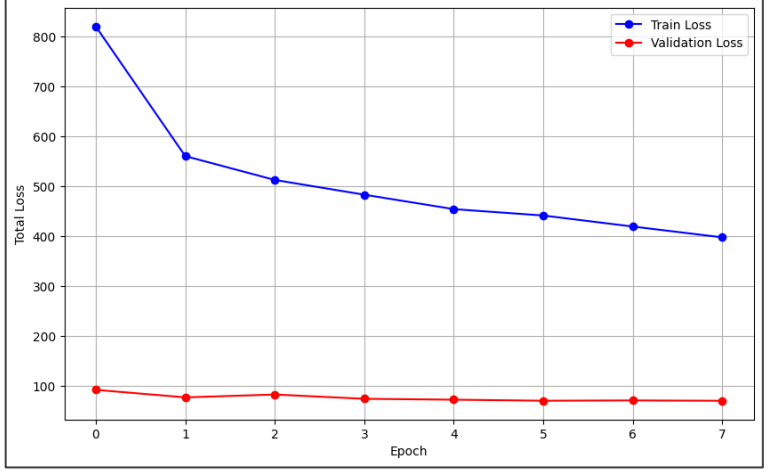
Использует все структуры

```
data range 6.7109952669829545 4.348977225287308e-05
line_graph True
100% 4768/4768 [00:00<00:00, 2807981.11it/s]
Reading dataset Atrain_data
data range 4.408488083342374 0.0
line_graph True
100% 596/596 [00:00<00:00, 2334085.14it/s]
Reading dataset Aval_data
data range 4.8731407918931 0.000225536621609
line_graph True
100% 596/596 [00:00<00:00, 2537873.28it/s]
Reading dataset Atest_data
n_train: 50
n_val  : 6
n_test : 6
rank 0
```

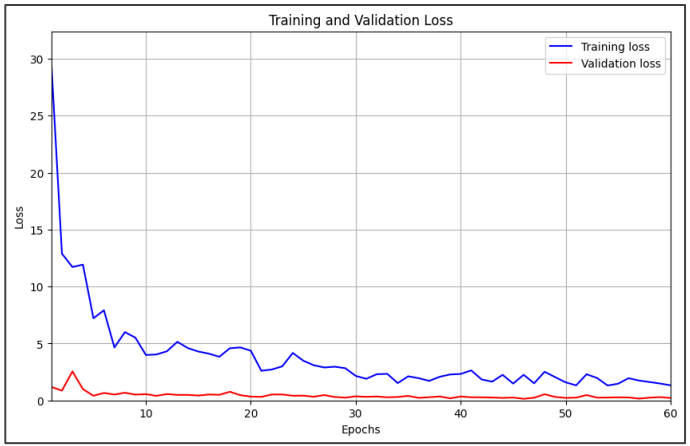
Использует не все структуры (очень мало)
Почему?

Проблемы

1. Не полная документация
2. Использует не все конфигурации (через раз)
3. Непонятно почему Loss для val такой маленький



Нормальное обучение



Не нормальное обучение (на малом количестве данных)

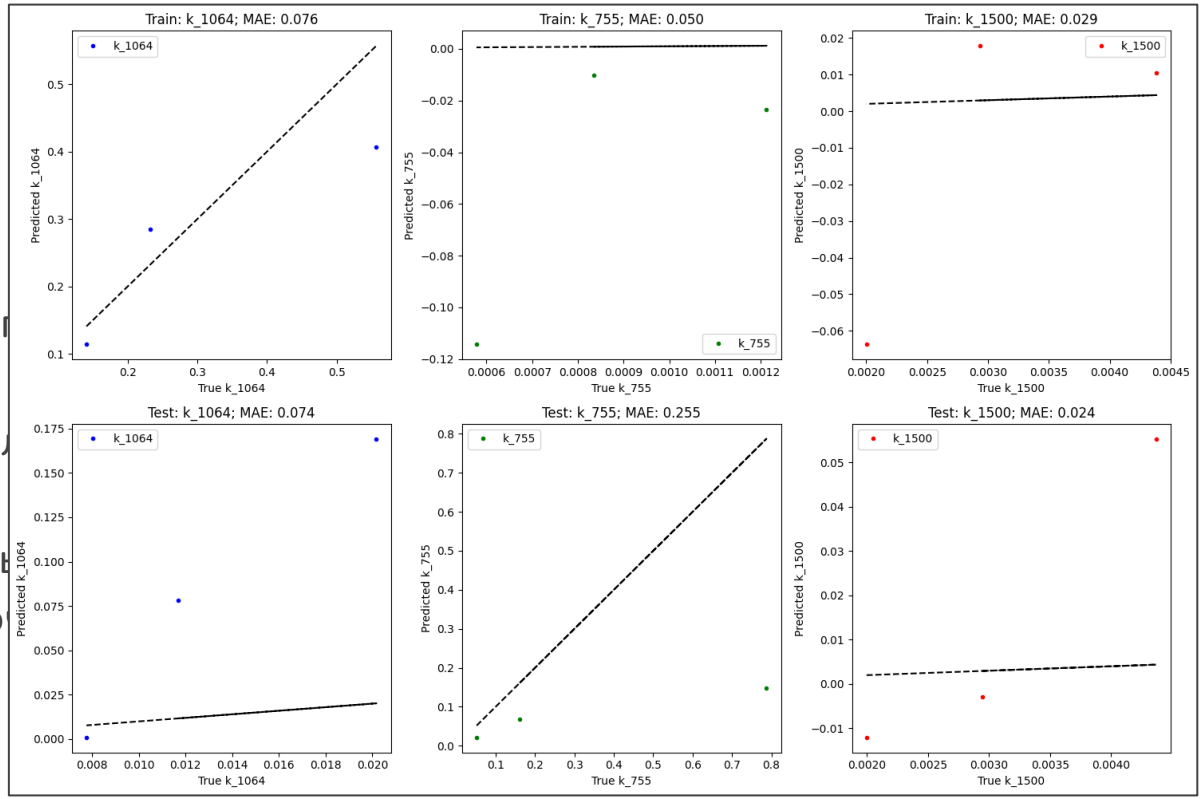
Проблемы

1. Не полная документация
2. Использует не все конфигурации (через раз)
3. Непонятно почему Loss для val такой маленький
4. В качестве результатов выводит рандомное количество точек (не разобрался до конца)

```
Overall MAE:
k_1064:
  Train MAE: 0.0764
num points: 3
  Test MAE:  0.0741
num points: 3
k_755:
  Train MAE: 0.0502
num points: 3
  Test MAE:  0.2548
num points: 3
k_1500:
  Train MAE: 0.0289
num points: 3
  Test MAE:  0.0236
num points: 3
```

Проблемы

1. Не полная документация
2. Использует не все конфи
   (через раз)
3. Непонятно почему Loss д
   маленький
4. В качестве результатов в
   рандомное количество то
   разобрался до конца)

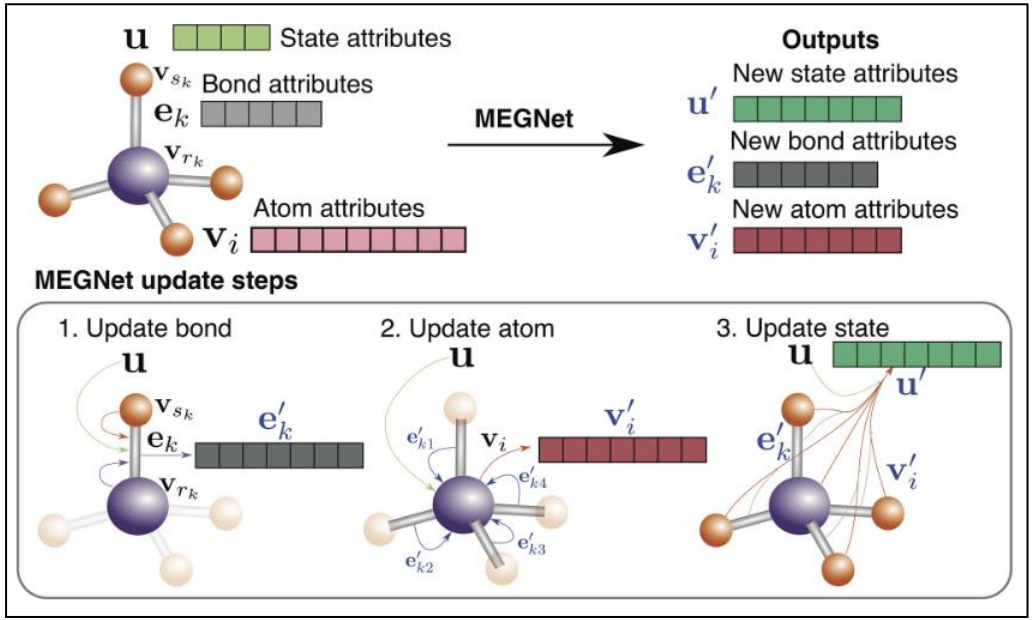


Еще GNN (Проблемы продолжаются)

crystal_example.ipynb

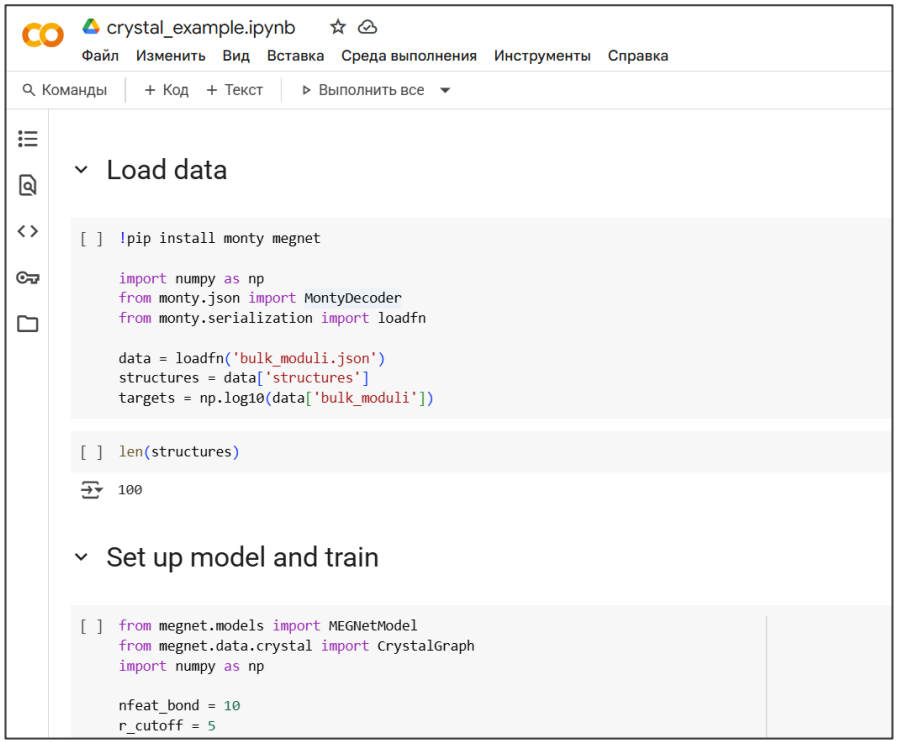
MegNet

1. Решено использовать другую модель



MegNet

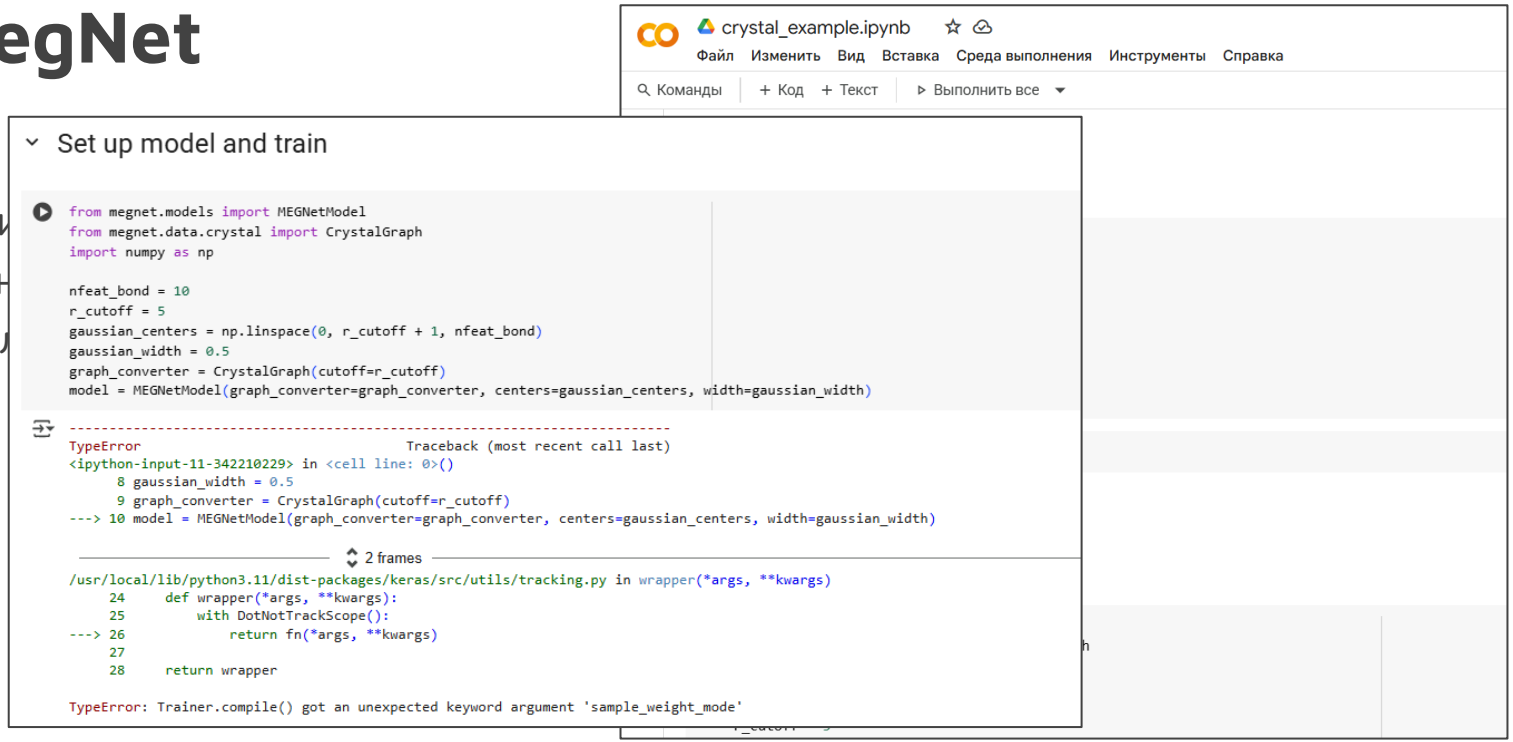
1. Решено использовать другую модель
2. Найден ноутбук в качестве примера (официальный - git)



https://colab.research.google.com/drive/1CHAsLh0klj5K82f8ZOcK9pLfV9Bx4OgK#scrollTo=A-LIUEDKIq1F

MegNet

1. Решено и
2. Найден н
   (официал



Не работает тестовый ноутбук?!

Возврат к истокам

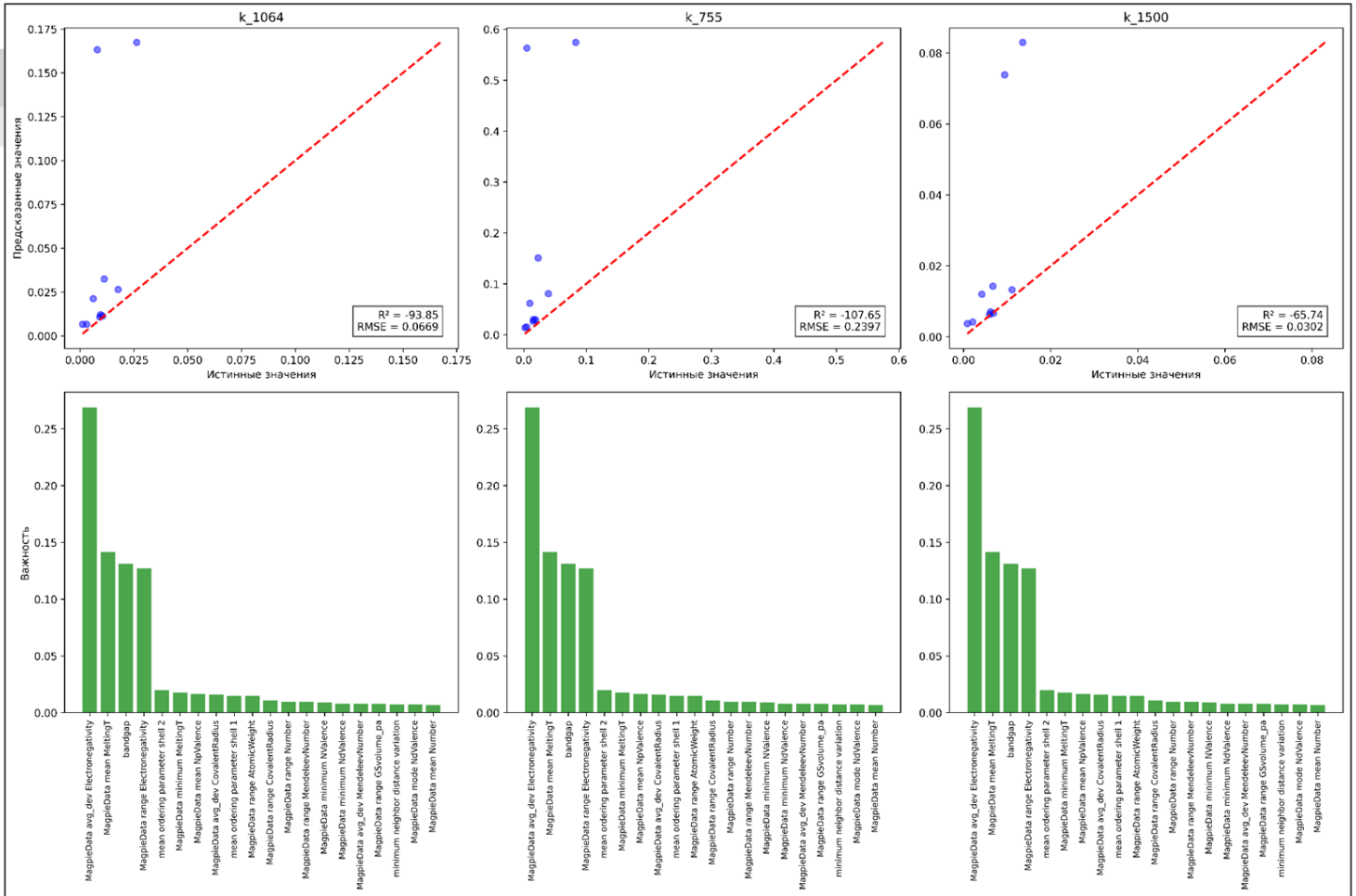
MultipleFeaturizerRegressor.ipynb

Структурные дескрипторы для классических регрессоров

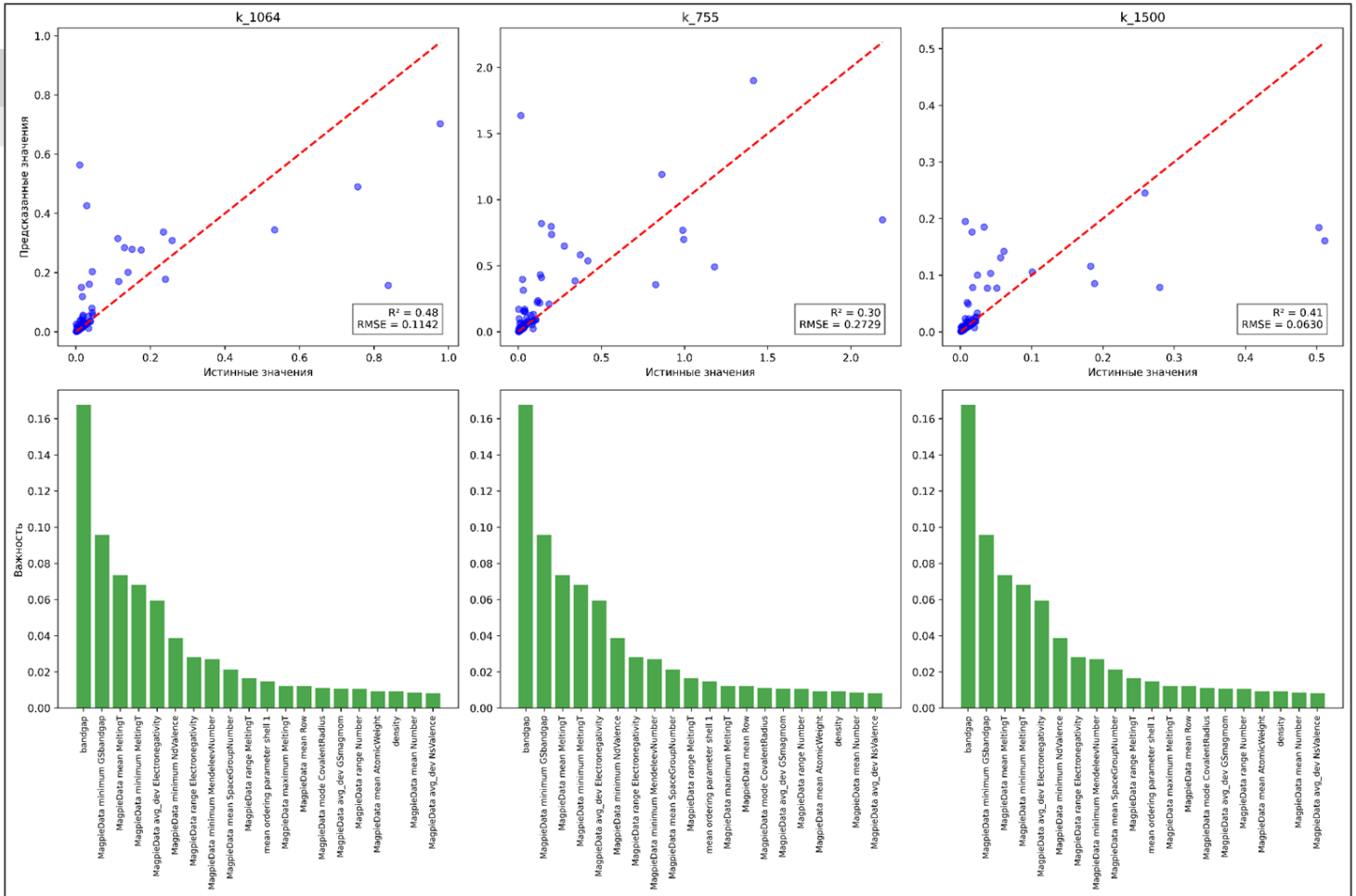
```
# Извлечения признаков из структуры
featurizer = MultipleFeaturizer([
    SiteStatsFingerprint.from_preset("CoordinationNumber_ward-prb-2017"),
    StructuralHeterogeneity(),
    ChemicalOrdering(),
    StructureComposition(ElementProperty.from_preset("magpie")),
    DensityFeatures()
])

X_features = featurizer.featurize_many(data["structure"], pbar=False, ignore_errors=True )
X_features = pd.DataFrame(X_features, columns=featurizer.feature_labels())
```

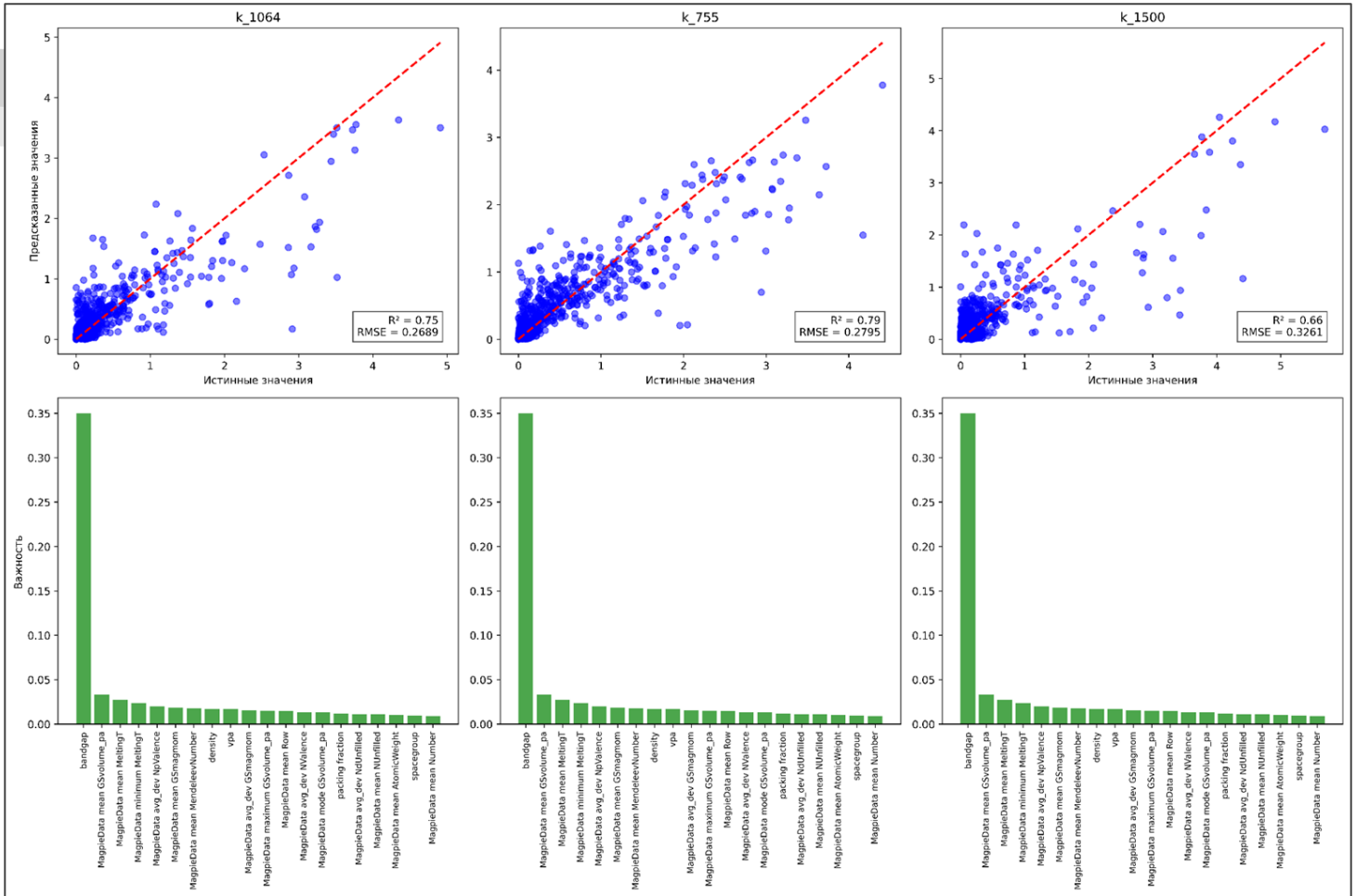
Результаты для разных размеров выборок (N=50)



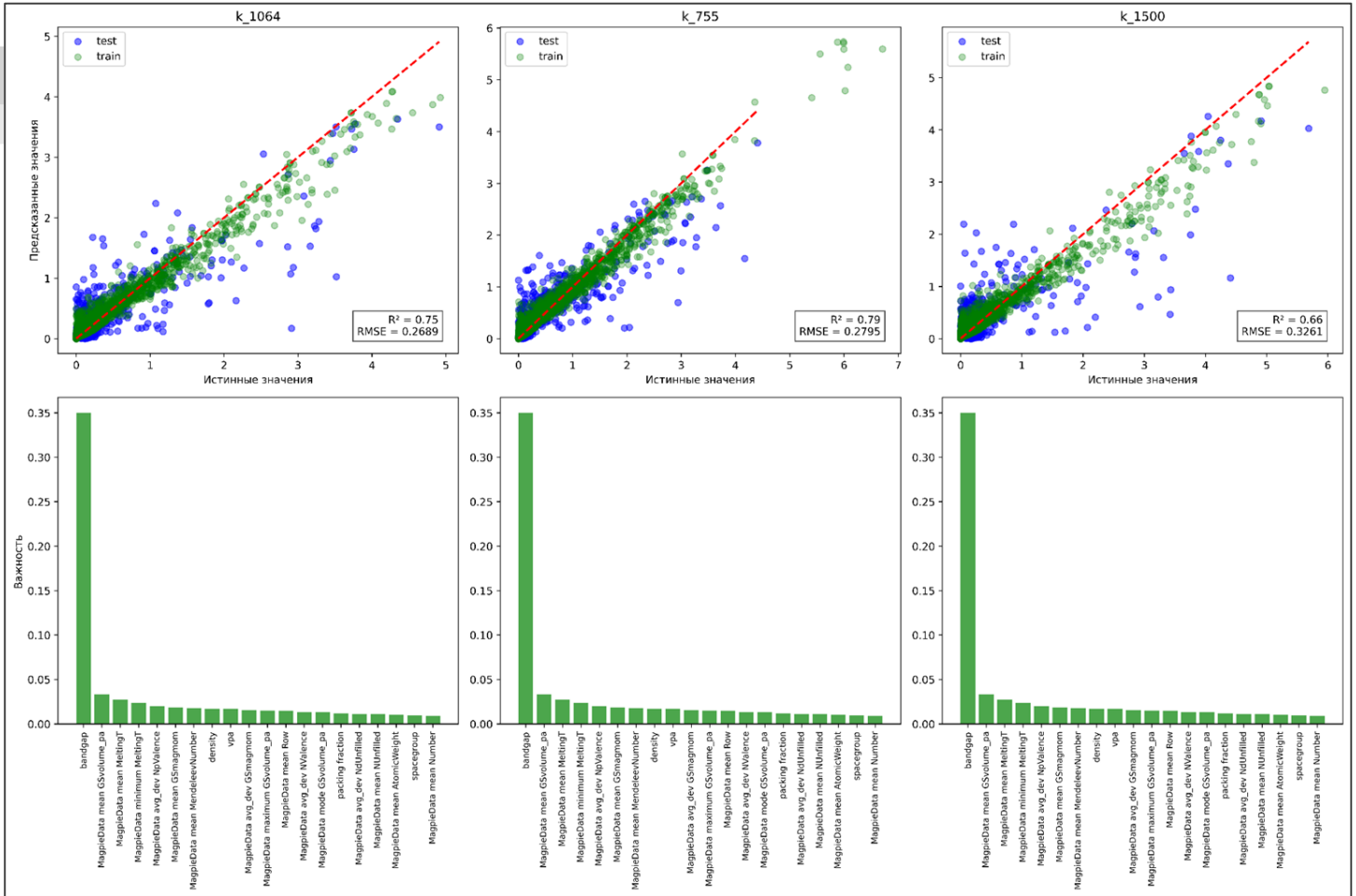
Результаты для разных размеров выборок (N=500)



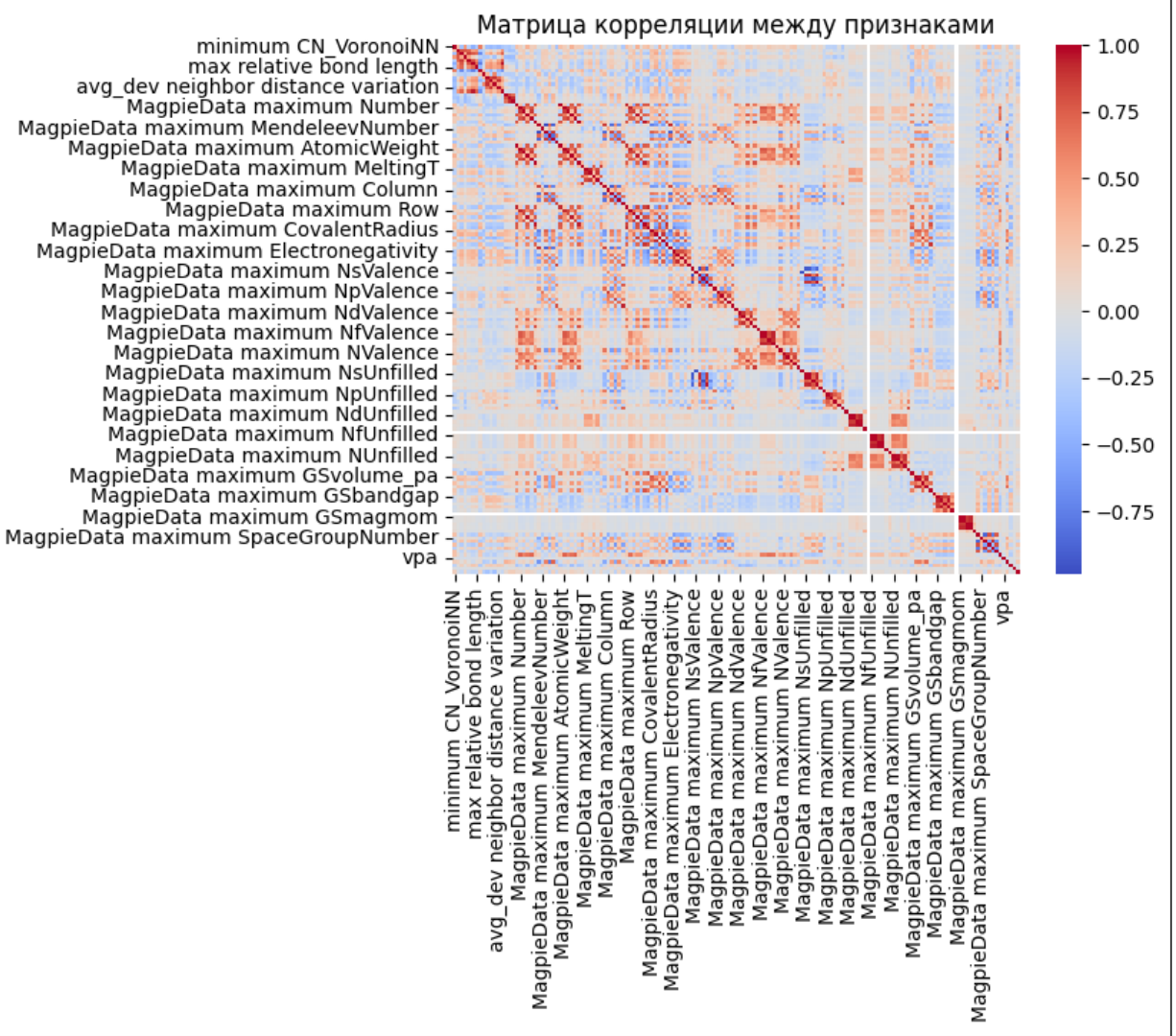
Результаты для разных размеров выборок (N=5960)



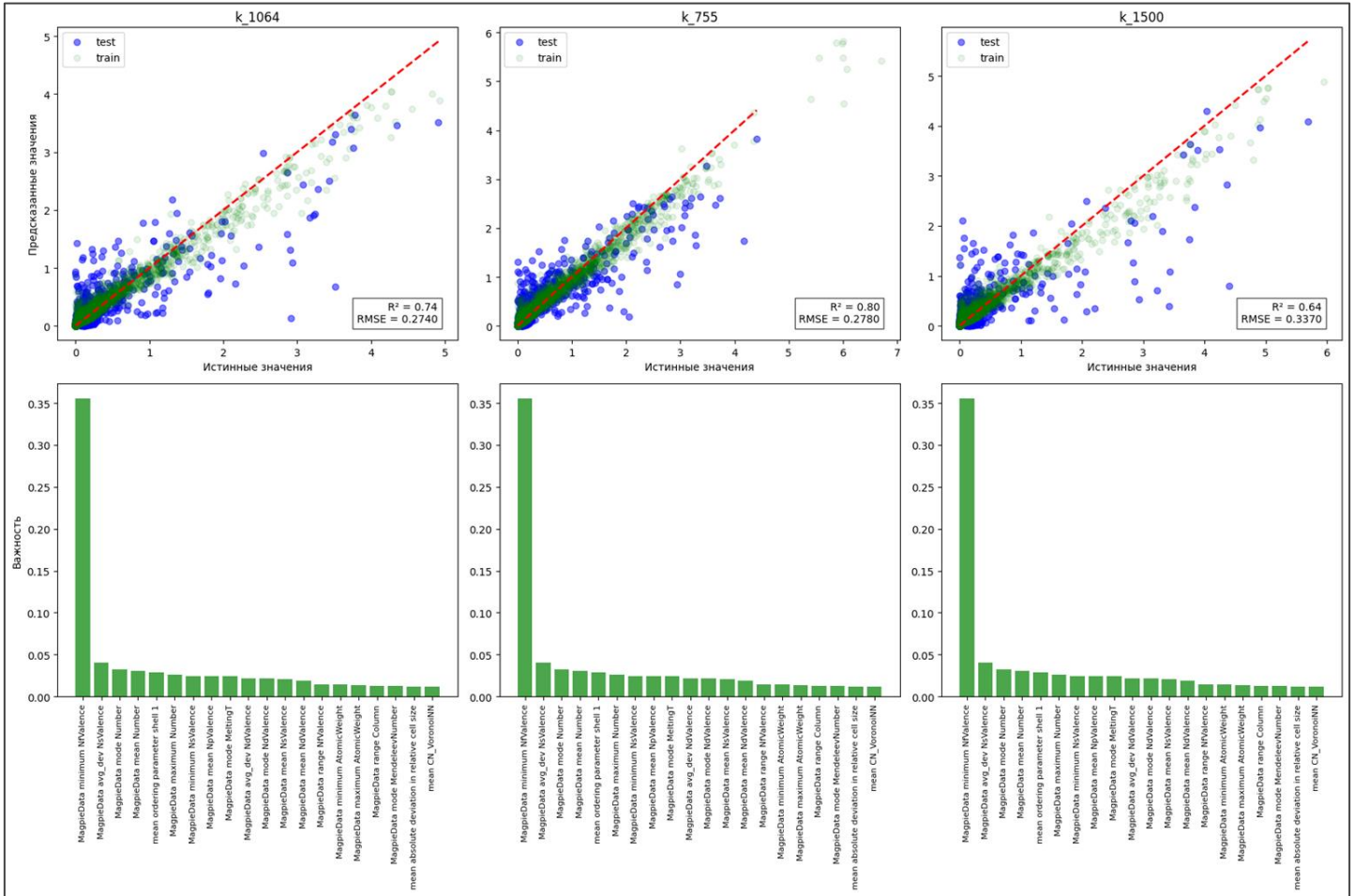
Результаты для GridSearch



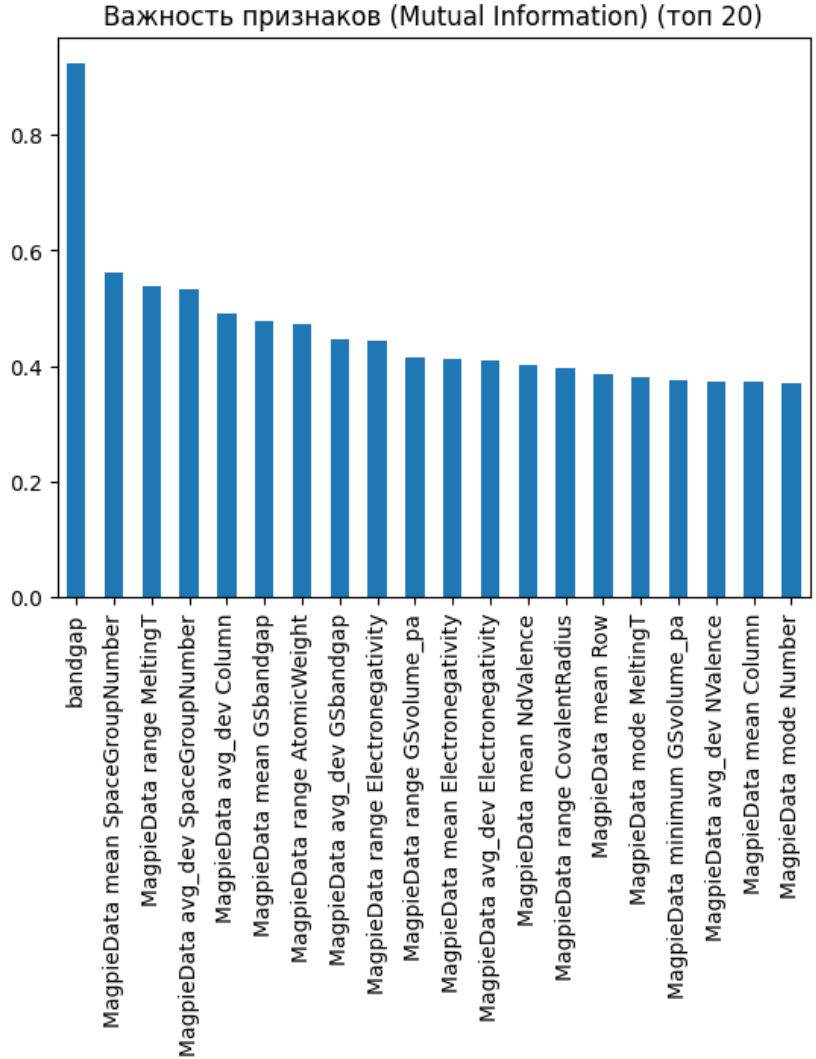
Корреляции


Матрица корреляции между признаками
minimum CN_VoronoiNN
max relative bond length
avg_dev neighbor distance variation
MagpieData maximum Number
MagpieData maximum MendeleevNumber
MagpieData maximum AtomicWeight
MagpieData maximum MeltingT
MagpieData maximum Column
MagpieData maximum Row
MagpieData maximum CovalentRadius
MagpieData maximum Electronegativity
MagpieData maximum NsValence
MagpieData maximum NpValence
MagpieData maximum NdValence
MagpieData maximum NfValence
MagpieData maximum NValence
MagpieData maximum NsUnfilled
MagpieData maximum NpUnfilled
MagpieData maximum NdUnfilled
MagpieData maximum NfUnfilled
MagpieData maximum NUnfilled
MagpieData maximum GSvolume_pa
MagpieData maximum GSbandgap
MagpieData maximum GSmagmom
MagpieData maximum SpaceGroupNumber
vpa
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

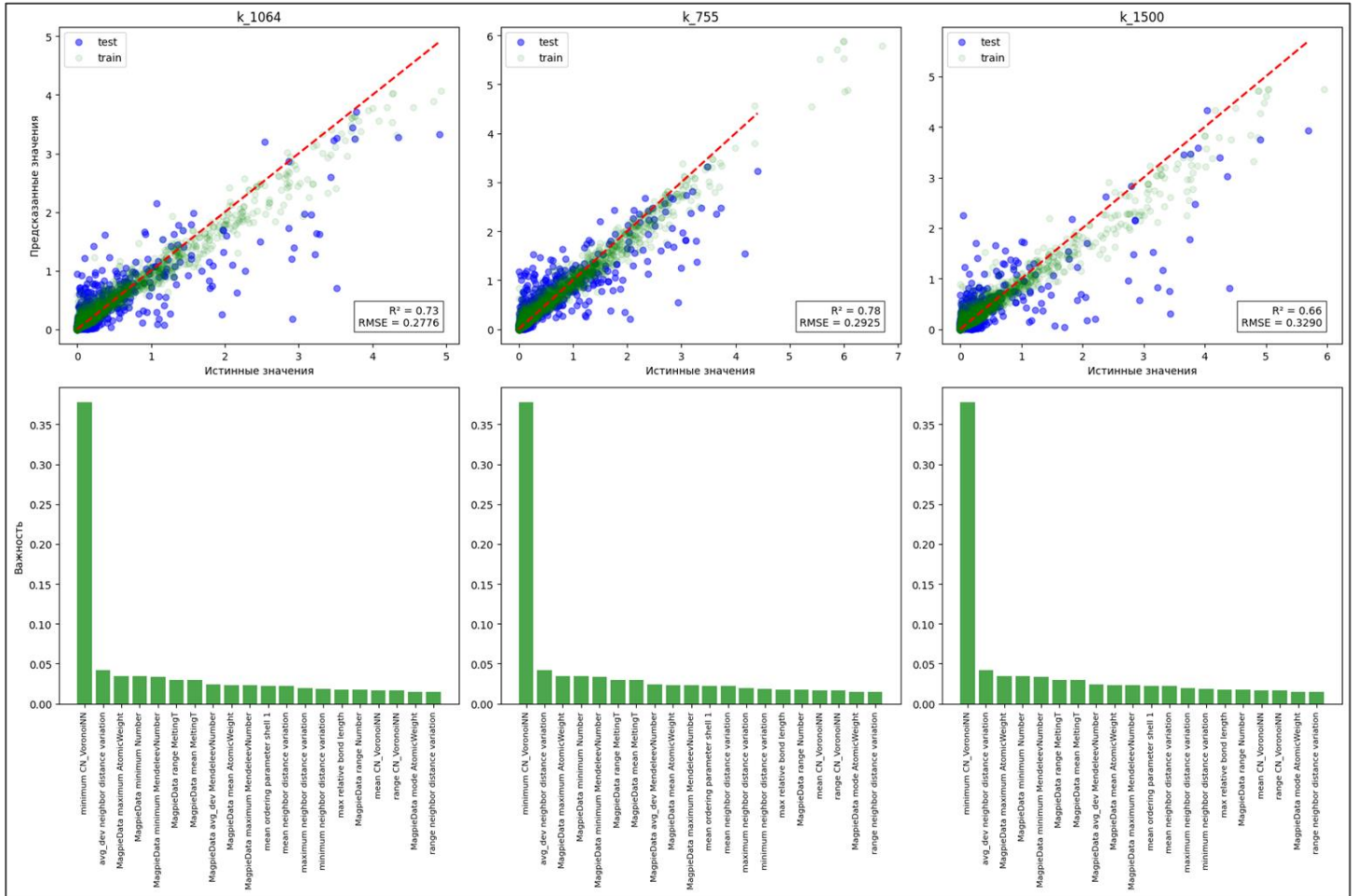
Результаты после удаления по корреляции



Нелинейный анализ



Результаты для топ40 по MI



Итоги

1. Обучено несколько моделей
2. На оптические свойства больше всего влияет band gap

@pechenkapop

vandyshev.gk@phystech.edu