Investigating the impact of descriptor generation on predictive modeling of optical properties for fluorescent molecules

Aisana Bolatbek

Nazarbayev University

Scientific Computing - Fall 2023
November 16, 2023

Presentation Overview

Introduction

Thesis Research Motivation:

- Fluorescent probes are essential tools in the fields of molecular biology, pharmacology, and cellular imaging.
- The accurate engineering of these probes is a difficult and time-consuming procedure that require numerous trials of experiments.
- Machine learning (ML) can greatly reduce computational costs, shorten the development cycle, and improve computational accuracy.

Thesis Research objectives:

- Harness predictive capabilities of machine learning to utilize a database of organic compounds.
- Leverage molecular representations and feature extraction to enhance usage of database.

Introduction

Observations:

- K-Nearest Neighbors Algorithm (KNN) *("Manhattan distance", weighted, K =5)* works better than other ML models
- There are a loads of descriptors. So the choice is another issue.
- There are different parameters set to generate descriptor. Dimension affects scores of predictive models

Objective

- Investigate the effect of descriptor generation on predictive modelling
- Assess computational cost of resulting model

Model Formulation

Algorithm of KNN Regression Model

1. Load the training and test data
2. Choose the value of K (K = 5)
3. For each point in test data:
   - find the Manhattan distance to all training data points
   - store the Manhattan distances in a list and sort it
   - choose the first 5 points
   - calculate the inverse distance weighted of the numerical target of first 5 points

Descriptor generation

1. Load the training and test data
2. Choose the descriptor (Morgan fingerprint as a bit array)
3. Choose the (bit) size of fingerprint
4. For each point in test data:
   - generate fingerprint as a bit array

Model parameters

Target Variable	Maximum Emission Wavelength
Independent Variables	Molecules of Chromophores and Solvents
Train / Test split	12698 / 2722
Regression Model	KNN
Model Parameters - Distance	$d(x,y) = \sum_{i=1}^{n} \lvert x_i - y_i \rvert$, where $n = length(x), x, y$ - data points
Model Parameters - Averaging	$\overline{x} = \frac{\sum_{i=1}^{5} w_i x_i}{\sum_{i=1}^{5} w_i}, w_i = \frac{1}{\sum_{j=1}^{5} d(x_i, x_j)}$, where $x_1, \ldots x_5$ - data points
Descriptor	Morgan Fingerprint
Descriptor length	$32, 64, 128, 256, 512, 1024$

Software and libraries

Python, Jupyter Notebook
Pandas, Numpy, Seaborn, Matplotlib
RDKit

Results Validation

Regression Evaluation Metric	$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \overline{y})^2}$, where y_i - true labels, f_i - predicted labels, $\overline{y}$ - mean of true values
Average Execution Time	$T = T_f + T_g + T_p$, where T_f - train data fit time, T_g - average descriptor generation time, T_p - average prediction time,

Validation Rule

if $R^2 > 0.8$ and $T < 1$, **then**
    Testing passed successfully,
else
    Testing failed.

Results

```
Testing 1 failed
R2 score = 0.6
Average execution time = 0.06061
-----
Testing 2 passed succesfully
R2 score = 0.81
Average execution time = 0.06678
-----
Testing 3 passed succesfully
R2 score = 0.84
Average execution time = 0.11
-----
Testing 4 passed succesfully
R2 score = 0.85
Average execution time = 0.22783
-----
Testing 5 passed succesfully
R2 score = 0.84
Average execution time = 0.43461
-----
Testing 6 failed
R2 score = 0.84
Average execution time = 1.07843
-----
```

Figure: Testing results

Results

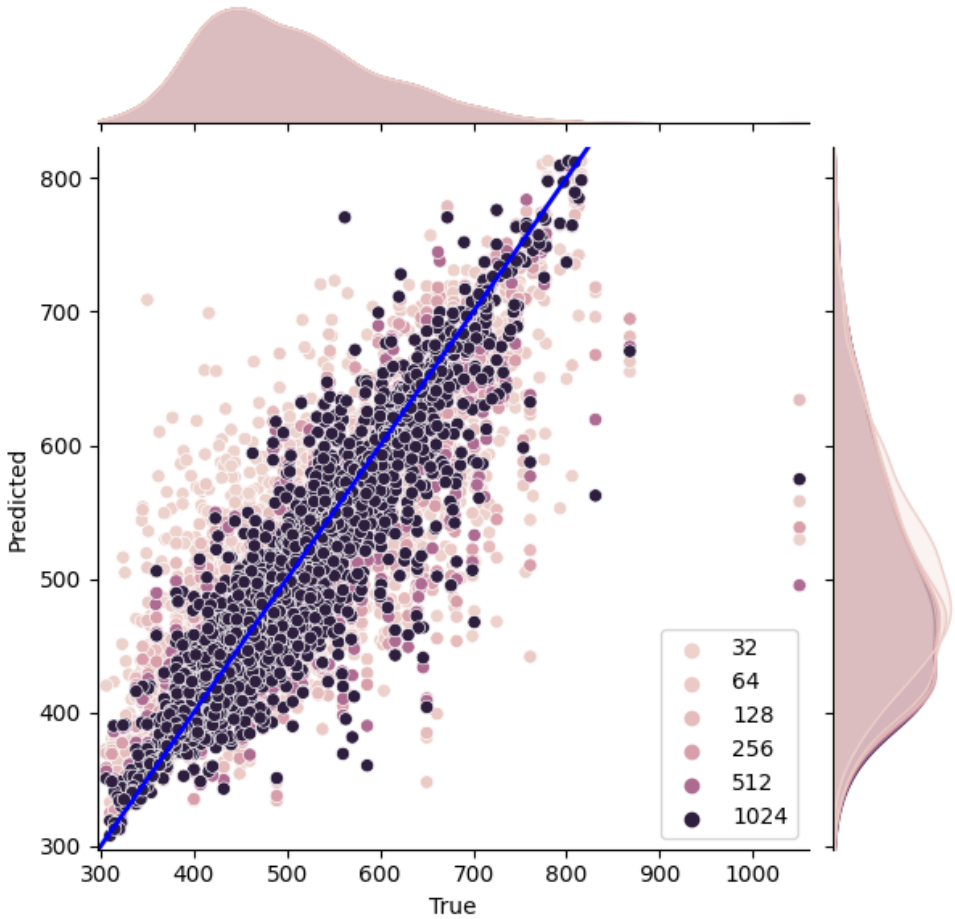

Figure: Predictions versus true values

Results

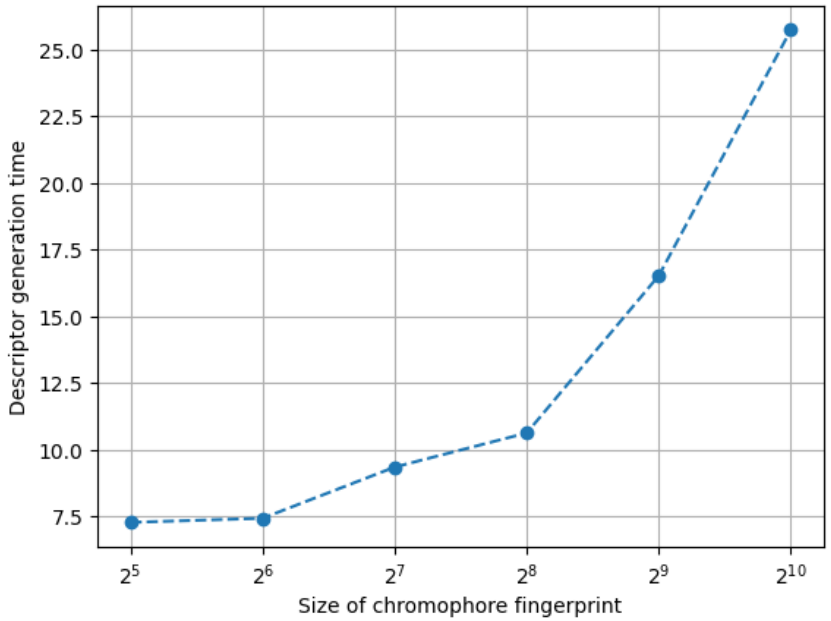

Figure: Descriptor generation time

Results

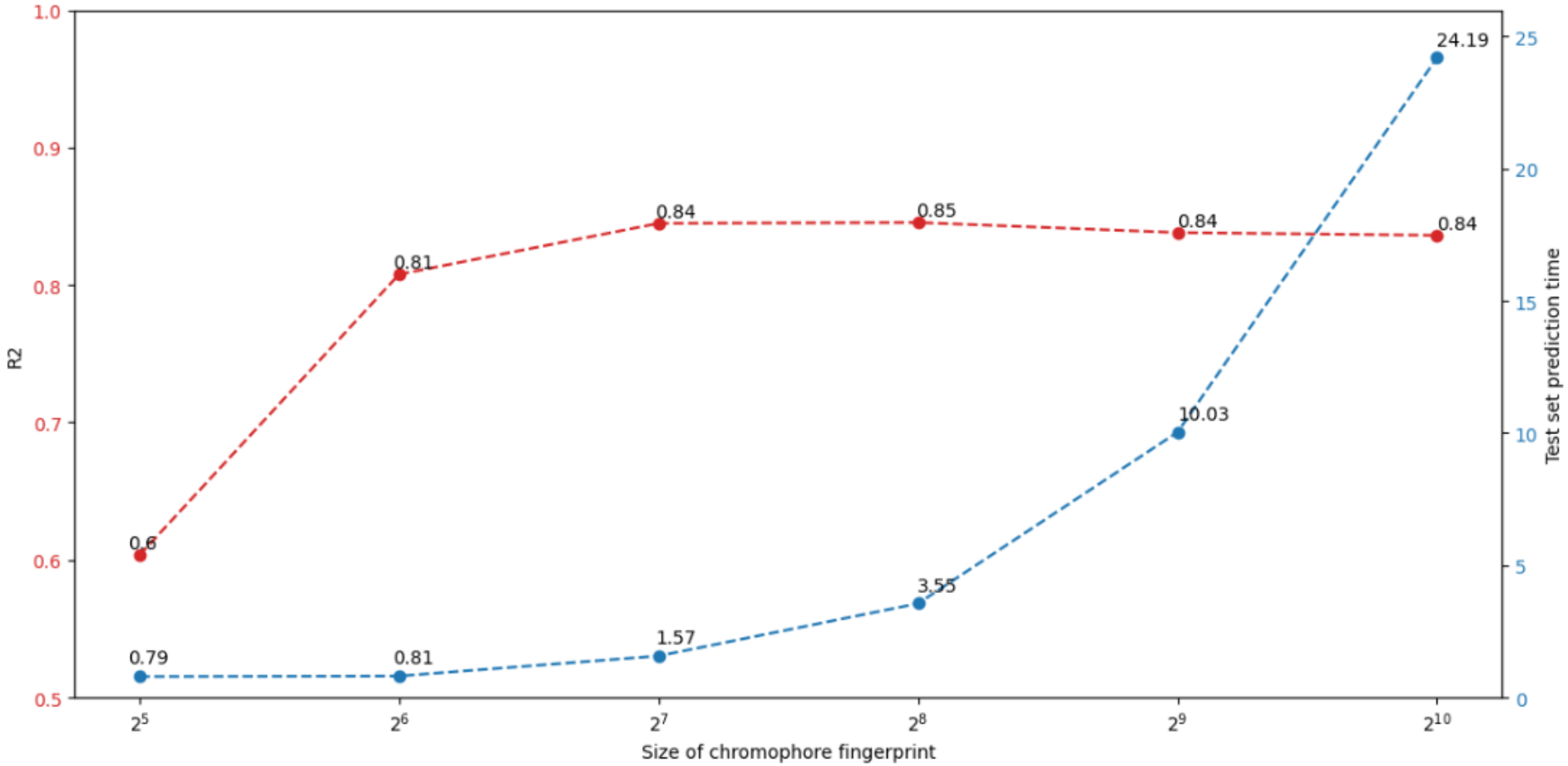

Figure: R2 scores versus Test set prediction time

Results

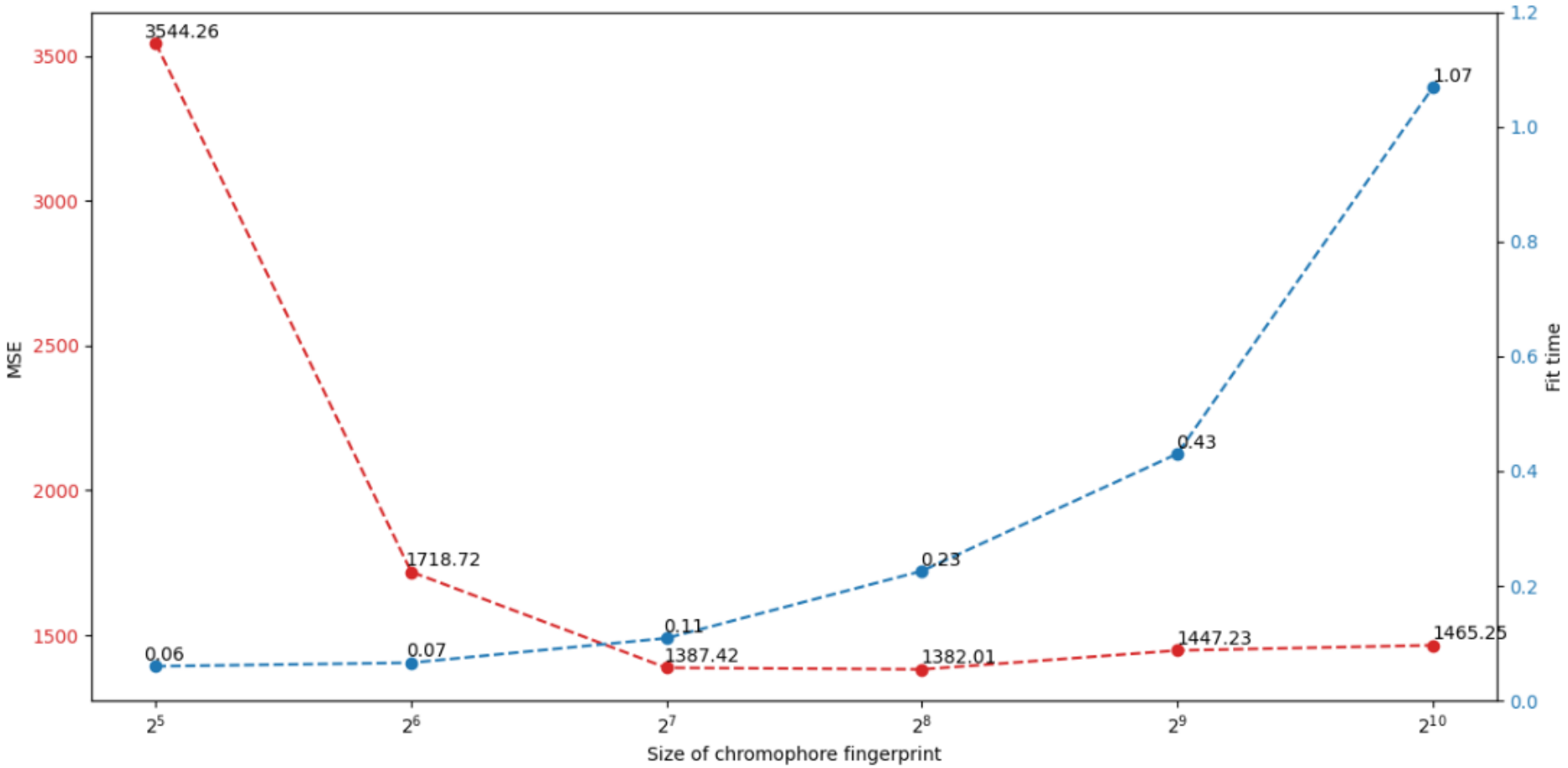

Figure: MSE scores versus Train set fit time

Version Control

Code is available at Github.

Dataset is presented in this paper.

Thank you !