# Explaining Diabetes Progression with BMI: Distributional Modeling, Robust Visualization, and Rigorous Regression Diagnostics

Gabrielprogramist

September 28, 2025

## 1 Problem Description

I investigate whether Body Mass Index (BMI)—measured at baseline—explains variation in one-year diabetes progression (the canonical quantitative target in the scikit-learn Diabetes dataset). I use distributional analysis (KDE with principled bandwidth selection), expressive visualization (X/Y position, color/markings, conditioning, and contextual overlays), and statistical modeling (from simple linear regression to multiple regression with robust inference and regularization).

## 2 Problem Details

- **Input features/signals:** standardized clinical baselines (age, bmi, bp, s1–s6). BMI is the primary explanatory variable; others are secondary controls.
- **Output:** predicted/expected diabetes progression (continuous target).
- **Objective:** quantify and validate the association between BMI and the target while avoiding misleading conclusions.

## 3 Methodology

### 3.1 Kernel Density Estimation (KDE)

The univariate Gaussian KDE for BMI is defined as:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left( \frac{x - x_i}{h} \right), \qquad K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}. \tag{1}$$

Bandwidth rules of thumb:

$$h_{\text{Scott}} = \sigma \, n^{-1/5}, \tag{2}$$

$$h_{\text{Silverman}} = 0.9 \min\left( \sigma, \frac{\text{IQR}}{1.34} \right) n^{-1/5}. \tag{3}$$

Cross-validation bandwidth is chosen by maximizing leave-one-out log-likelihood:

$$\ell(h) = \sum_{i=1}^{n} \log \hat{f}_{-i,h}(x_i), \quad \hat{f}_{-i,h}(x_i) = \frac{1}{(n-1)h} \sum_{j \neq i} K\left(\frac{x_i - x_j}{h}\right). \tag{4}$$

### 3.2 Simple Linear Regression (SLR)

$$y = \beta_0 + \beta_1 \,\mathrm{BMI} + \varepsilon, \qquad \hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{5}$$

Evaluation metric:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}. \tag{6}$$

Breusch–Pagan test checks for heteroskedasticity:

$$\mathrm{BP} \sim \chi^2_p, \tag{7}$$

from regression of $e_i^2$ on predictors.

Cook's distance for influential observations:

$$D_i = \frac{e_i^2}{p\hat{\sigma}^2} \cdot \frac{h_{ii}}{(1 - h_{ii})^2}. \tag{8}$$

### 3.3 Beyond SLR

- Polynomial regression in BMI (degree $d$ selected by CV).
- Multiple regression with robust HC3 standard errors for BMI's partial effect.
- Regularization methods:

$$\text{Ridge:} \quad \min_{\beta} \frac{1}{n} \sum (y_i - \mathbf{x}_i^\top \beta)^2 + \lambda \|\beta\|_2^2, \tag{9}$$

$$\text{Lasso:} \quad \min_{\beta} \frac{1}{n} \sum (y_i - \mathbf{x}_i^\top \beta)^2 + \lambda \|\beta\|_1. \tag{10}$$

## 4 Data

- **Diabetes dataset (scikit-learn):** 442 rows, 10 standardized predictors, target is 1-year diabetes progression.

Table 1: BMI KDE bandwidths: rules of thumb vs CV-optimal.

| Bandwidth | Scott | Silverman | CV-optimal |
|---|---|---|---|
| $h$ | 0.014 | 0.013 | 0.011 |

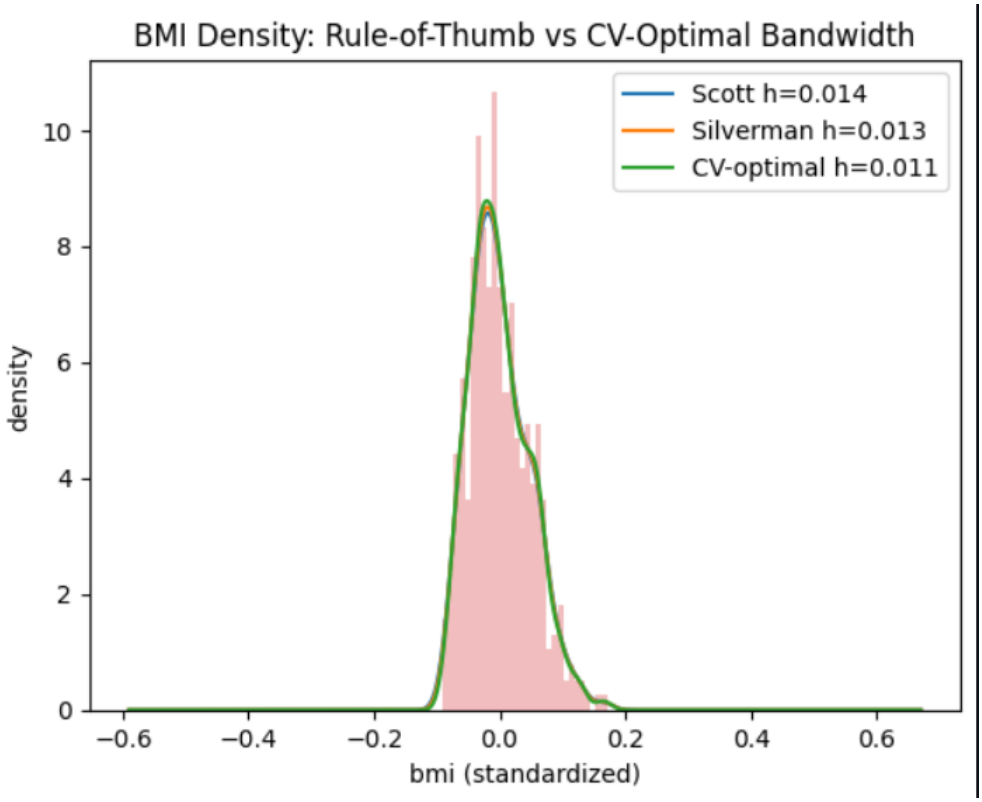


Figure 1: BMI density with Scott, Silverman, and CV-optimal bandwidths. Curves are very similar; CV is slightly narrower ($\approx 0.011$).

Table 2: SLR summary (bmi-only).

| Metric | Value |
|---|---|
| Slope (bmi) | 949.4353 |
| Intercept | 152.1335 |
| RMSE | 62.3735 |

Table 3: Test for heteroskedasticity in SLR residuals.

| Breusch–Pagan p-value: 0.000853 |
|---|

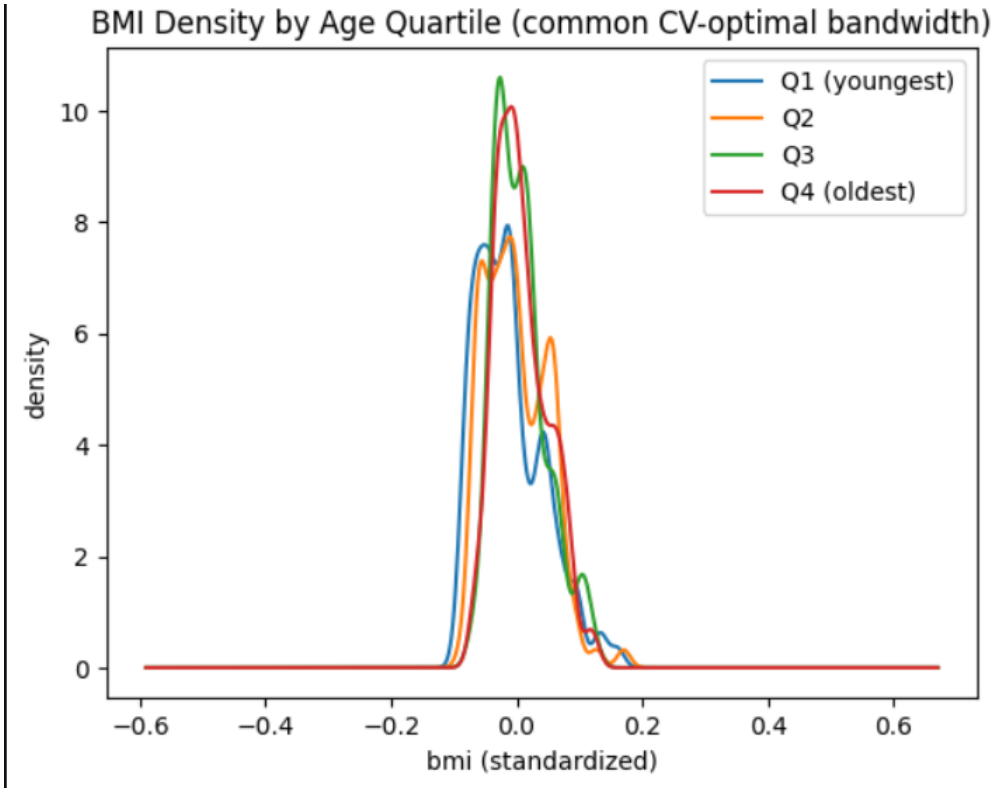


Figure 2: BMI density by age quartile using a common CV-optimal bandwidth. Central mass is similar; tails diverge modestly.

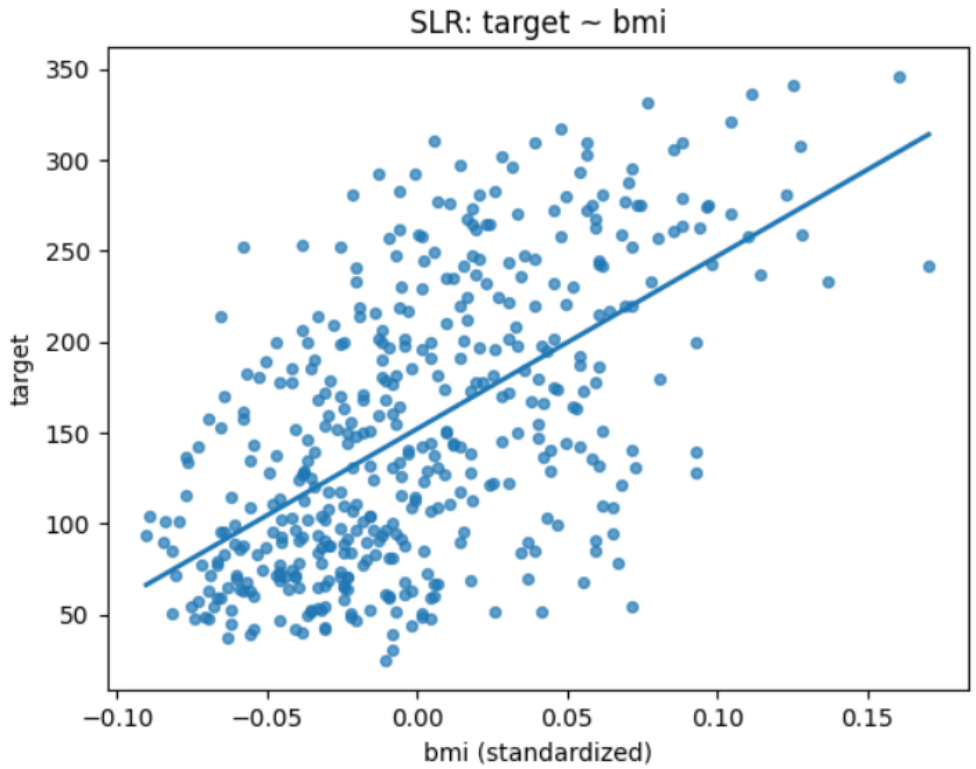


Figure 3: Simple linear regression: `target` vs. `bmi`. Clear positive association.

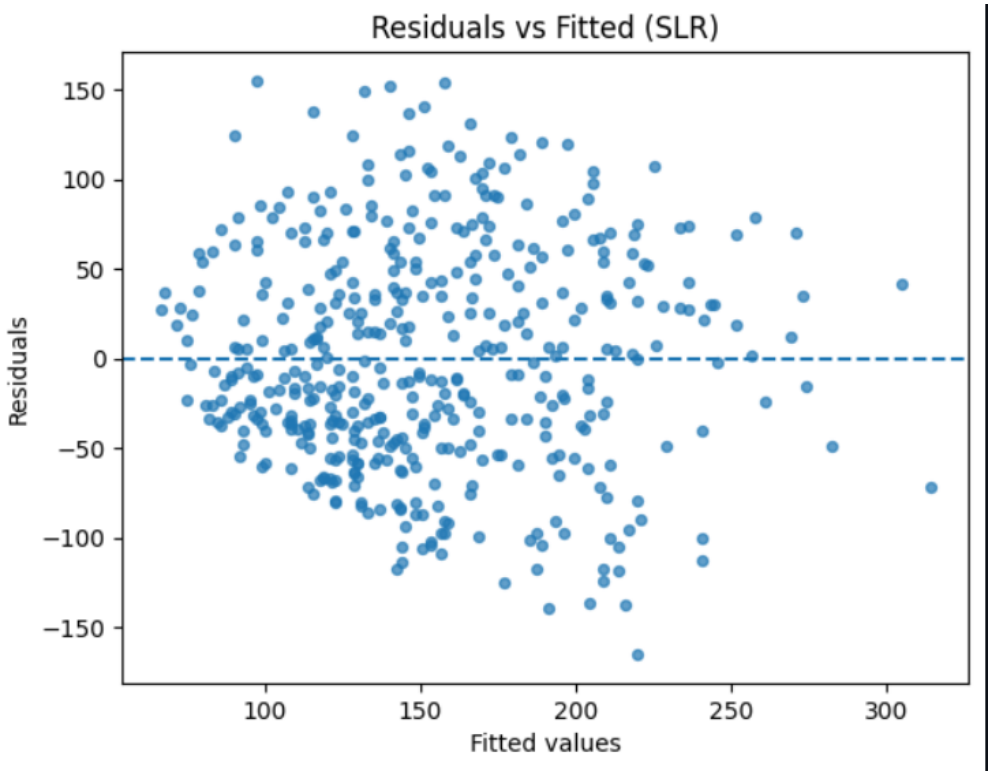


Figure 4: Residuals vs fitted for SLR: heteroskedasticity pattern (funnel-like spread).

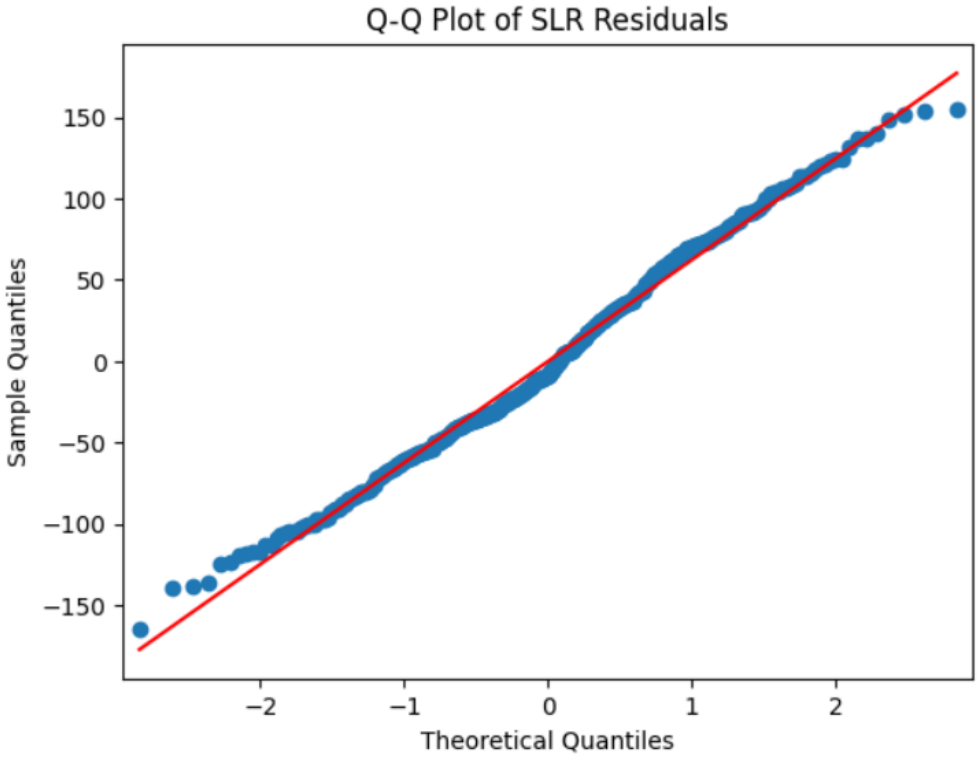


Figure 5: Q–Q plot for SLR residuals: near-linear with mild tail deviations.

Table 4: Polynomial degree selection for BMI (10-fold CV).

| Degree | CV_RMSE |
|---|---|
| 1 | 62.604289 |
| 2 | 62.799326 |
| 3 | 62.989974 |
| 4 | 63.332757 |
| 5 | 62.911145 |

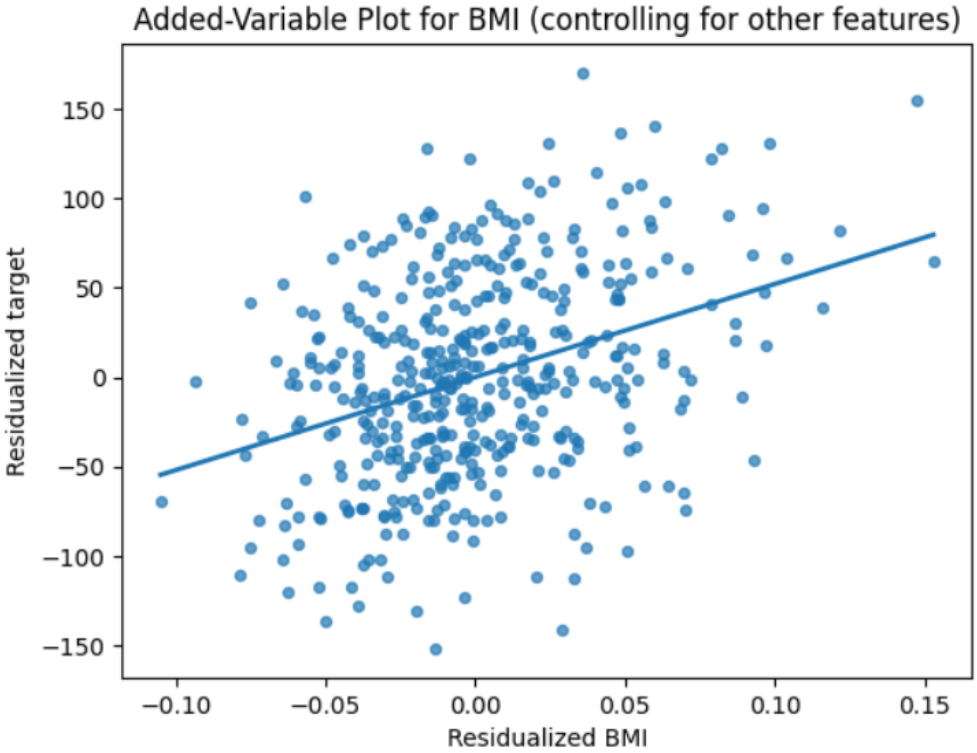


Figure 6: Added-variable plot: partial effect of `bmi` controlling for other features. Positive partial trend.

Table 5: Multiple regression: partial effect of `bmi` with HC3 standard errors.

| Term | Coef | Robust SE (HC3) | t-value |
|---|---|---|---|
| bmi (partial) | 519.8459 | 68.5908 | 7.5790 |

Table 6: Model comparison by 10-fold CV RMSE. Lower is better.

| Model | CV_RMSE |
|---|---|
| Lasso (all features) | 54.837533 |
| Full OLS (all features) | 54.860615 |
| Ridge (all features) | 55.042534 |
| SLR (BMI only) | 62.604289 |
| Polynomial BMI (deg=1) | 62.604289 |

Table 7: BMI coefficient across models (stability under regularization).

| Feature | OLS_full | Ridge | Lasso |
|---|---|---|---|
| bmi | 519.84592 | 521.05693 | 521.172424 |

# 5 Results

## 5.1 Exploratory Distributional Analysis

## 5.2 SLR Fit and Diagnostics

## 5.3 Beyond SLR: Nonlinearity and Controls

## 5.4 Regularization and Model Comparison

## 5.5 Interpretation Summary

- **BMI distribution.** Scott $\approx 0.014$, Silverman $\approx 0.013$, CV-optimal $\approx 0.011$. Curves are very similar; subgroup KDEs by age show similar centers with modest tail differences.
- **SLR.** Clear positive slope (949.4); RMSE $\approx 62.37$. Residuals exhibit heteroskedasticity (Breusch–Pagan $p = 0.000853$); QQ-plot is near-linear with mild tail deviations.
- **Nonlinearity.** 10-fold CV selects degree = 1; higher degrees do not improve RMSE $\Rightarrow$ linear effect in BMI is adequate at this sample size.
- **Controls.** In multiple regression with HC3, BMI remains strongly significant (coef $\approx 519.85$, $t \approx 7.58$); added-variable plot visualizes the positive partial effect.
- **Generalization.** Multivariate models materially outperform BMI-only (CV-RMSE $\approx 55$ vs 62.6). Ridge/Lasso give similar accuracy; BMI coefficient is stable across OLS/Ridge/Lasso ($\approx 520$).

# 6 Conclusion

- BMI is a robust positive predictor of diabetes progression.
- Heteroskedasticity in SLR requires robust inference (HC3) for valid uncertainty quantification.
- No strong nonlinearity in BMI's effect is detected; multivariate models substantially improve predictive accuracy.
- KDE with cross-validated bandwidth provides reliable distributional summaries and supports the modeling choices.

# 7 Data Source

- Diabetes dataset: `sklearn.datasets.load_diabetes()`.

# Appendix

## A. Top-10 Influential Points (Cook's Distance)

Table 8: Top-10 by Cook's distance (SLR).

| Cook's distance | Leverage |
|---|---|
| 0.025992 | 0.007360 |
| 0.022201 | 0.031352 |
| 0.018009 | 0.010903 |
| 0.017441 | 0.005620 |
| 0.016661 | 0.006763 |
| 0.014368 | 0.010903 |
| 0.013066 | 0.006551 |
| 0.012813 | 0.005313 |
| 0.012139 | 0.008158 |
| 0.011847 | 0.006478 |

## B. Full Coefficient Table (OLS vs Ridge vs Lasso)

Table 9: Coefficient comparison across models (all features).

| Feature | OLS_full | Ridge | Lasso |
|---|---|---|---|
| s1 | -792.175639 | -547.475235 | -92.831552 |
| s5 | 751.273700 | 657.788378 | 508.078335 |
| bmi | 519.845920 | 521.056930 | 521.172424 |
| s2 | 476.739021 | 282.626597 | -0.000000 |
| bp | 324.384646 | 322.467922 | 292.377164 |
| sex | -239.815644 | -237.277928 | -188.585504 |
| s4 | 177.063238 | 148.328962 | 0.000000 |
| s3 | 101.043268 | -6.507128 | -220.945511 |
| s6 | 67.626692 | 69.341322 | 50.210204 |
| age | -10.009866 | -8.509261 | -0.000000 |

## C. Full KDE Bandwidth Grid (LOO Log-Likelihood)

Table 10: KDE bandwidth grid for BMI: $h$ vs LOO log-likelihood.

| $h$ | $\ell(h)$ |
|---|---|
| 0.00253 | 707.239095 |
| 0.00276 | 711.468454 |
| 0.00301 | 715.071307 |
| 0.00327 | 718.027459 |
| 0.00356 | 720.635326 |
| 0.00388 | 722.898715 |
| 0.00423 | 724.838718 |
| 0.00460 | 726.448127 |
| 0.00501 | 727.847534 |
| 0.00546 | 729.045951 |
| 0.00594 | 730.038880 |
| 0.00647 | 730.880261 |
| 0.00705 | 731.568694 |
| 0.00768 | 732.106398 |
| 0.00836 | 732.499141 |
| 0.00910 | 732.758181 |
| 0.00991 | 732.889610 |
| 0.01080 | 732.895382 |
| 0.01176 | 732.776310 |
| 0.01280 | 732.529465 |
| 0.01394 | 732.138700 |
| 0.01518 | 731.583976 |
| 0.01653 | 730.833621 |
| 0.01800 | 729.846789 |
| 0.01961 | 728.565896 |
| 0.02135 | 726.950680 |
| 0.02325 | 724.919876 |
| 0.02532 | 722.400241 |
| 0.02757 | 719.312540 |
| 0.03003 | 715.540465 |
| 0.03270 | 711.004674 |
| 0.03561 | 705.572572 |
| 0.03877 | 699.141411 |
| 0.04222 | 691.544248 |
| 0.04598 | 682.652876 |
| 0.05007 | 672.349610 |
| 0.05453 | 660.481478 |
| 0.05938 | 646.971209 |
| 0.06466 | 631.721282 |
| 0.07042 | 614.642726 |