# Error-informed Gaussian process regression for predicting material science quantities

Anna Paulish
Supervisor: Prof. Michael Herbst

Mathematics for materials modelling group

March 28, 2025

MARVEL
NATIONAL CENTRE OF COMPETENCE IN RESEARCH

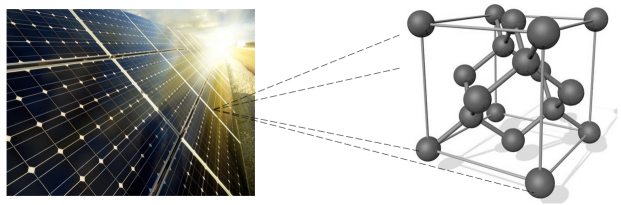
EPFL

MX t Mat

# Overview

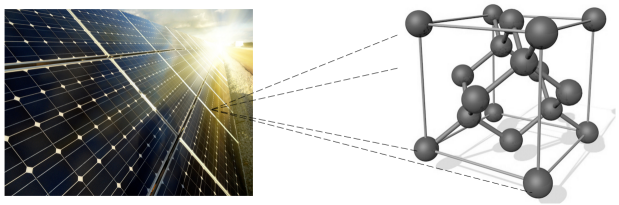
1. Introduction
   - Motivation: challenge of data generation
   - Background: Gaussian process regression for materials
2. Research goal: develop an error-informed data-driven model for predicting material properties
3. Results
   - Proof of principle: numerical integration
   - Application to realistic system: equation of state prediction
   - Error-informed Machine Learned Interatomic Potential (MLIP)
4. Conclusion and outline

# Material discovery and design

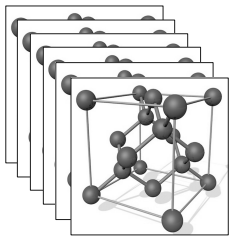


- Cornerstone of modern technology
- Goal: finding new materials with specific target properties
- Traditional way: Density Functional Theory (DFT) based **high-throughput computational screening**

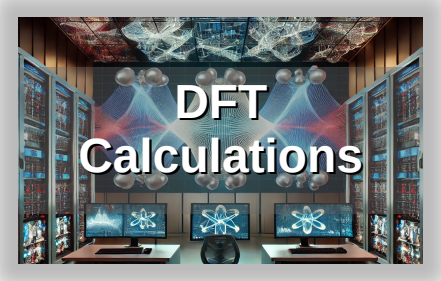
# Material discovery and design

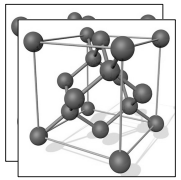


- Cornerstone of modern technology
- Goal: finding new materials with specific target properties
- Traditional way: Density Functional Theory (DFT) based **high-throughput computational screening**



Lots of structures



Take the best

✘ One DFT calculation can take a few **hours**

✘ Need **thousands**

# Data-driven acceleration


- Successful examples: GAP[1], SA-GPR[2], MACE[3], PET[4], MatterSim[5]
- Predictions for new structures take less than a **second**
- The model has to be trained $\rightarrow$ need to generate **training data**

# Data-driven acceleration


- Successful examples: GAP[1], SA-GPR[2], MACE[3], PET[4], MatterSim[5]
- Predictions for new structures take less than a **second**
- The model has to be trained → need to generate **training data**

Data generation:

- **Input**: structures → **DFT** → **Output**: materials properties
- Training data: {structures, DFT Outputs}

[1] Machine Learning a General-Purpose Interatomic Potential for Silicon. A. P. Bartók et al., Phys. Rev. X, 8, 041048, 2018
[2] Transferable Machine-Learning Model of the Electron Density. A. Grisafi et al., ACS Cent. Sci., 5, 57-64, 2018
[3] MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. I. Batatia et al., NeurIPS 2022
[4] Smooth, exact rotational symmetrization for deep learning on point clouds. S. N. Pozdnyakov and M. Ceriotti, NeurIPS 2023
[5] MatterSim: A Deep Learning Atomistic Model Across Elements, Temperatures, and Pressures. Han Yang et al., arXiv preprint arXiv:2405.04967, 2024

# Data generation process

**Common atomistic ML models** assume error in training data as small Gaussian noise:

$$y_i = f(x_i) + \epsilon_i, \text{ where } \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2),$$

where $i \in \{1, \ldots, n\}$, $n$ – number of training samples.
Homoscedastic noise covariance: $\Sigma = \sigma I$

# Data generation process

**Common atomistic ML models** assume error in training data as small Gaussian noise:

$$y_i = f(x_i) + \epsilon_i, \text{ where } \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2),$$

where $i \in \{1, \ldots, n\}$, $n$ – number of training samples.
Homoscedastic noise covariance: $\Sigma = \sigma I$

**Problem**

✘ All conducted simulations have to be homogeneously high quality.
✘ This is often infeasible for large systems.
✘ Data from different sources.

# How to incorporate error information into the model?

**Solution**

Error-informed GP regression:

$$y_i = f(x_i) + \epsilon_i, \text{ where } \epsilon_i \sim \mathcal{N}(0, \sigma_i^2).$$

# How to incorporate error information into the model?

**Solution**

Error-informed GP regression:

$$y_i = f(x_i) + \epsilon_i, \text{ where } \epsilon_i \sim \mathcal{N}(0, \sigma_i^2).$$

Heteroscedastic noise covariance: $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$

# How to incorporate error information into the model?

**Solution**

Error-informed GP regression:

$$y_i = f(x_i) + \epsilon_i, \text{ where } \epsilon_i \sim \mathcal{N}(0, \sigma_i^2).$$

Heteroscedastic noise covariance: $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$

✓ Reduce data generation cost while maintaining prediction accuracy.

✓ Discretization errors $\epsilon_i$ can be estimated[1].

[1] Eric Cancès, Geneviève Dusson, Gaspard Kemlin, and Antoine Levitt. Practical error bounds for properties in plane-wave electronic structure calculations. SIAM Journal on Scientific Computing, 44(5):B1312–B1340, 2022.

# Gaussian process regression (GPR)

**Gaussian process** is a collection of random variables, any finite number of which have a joint Gaussian distribution.

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

# Gaussian process regression (GPR)

**Gaussian process** is a collection of random variables, any finite number of which have a joint Gaussian distribution.

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

- Joint prior distribution of train $\mathbf{y}$ and test outputs $\mathbf{f}_* = f(x_*)$:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} m(X) \\ m(X_*) \end{bmatrix}, \begin{bmatrix} K(X,X) + \Sigma & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix} \right).$$

- Noise covariance: $\Sigma = \sigma^2 I$ vs $\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$

# Gaussian process regression (GPR)

**Gaussian process** is a collection of random variables, any finite number of which have a joint Gaussian distribution.

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

- Joint prior distribution of train $\mathbf{y}$ and test outputs $\mathbf{f}_* = f(x_*)$:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} m(X) \\ m(X_*) \end{bmatrix}, \begin{bmatrix} K(X,X) + \Sigma & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix} \right).$$

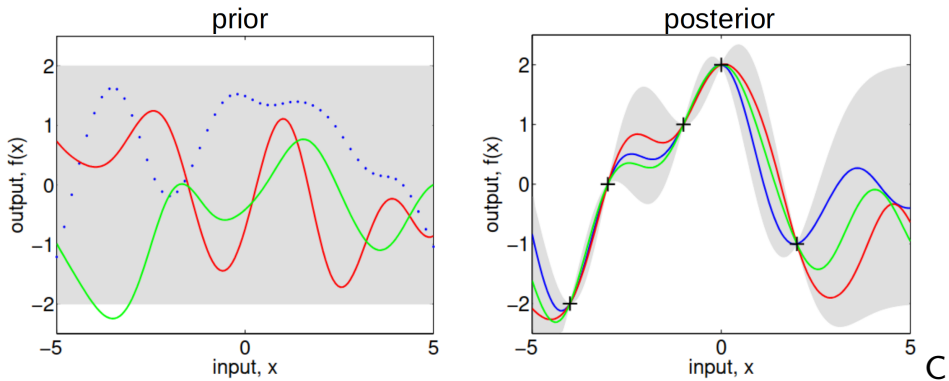
- Noise covariance: $\Sigma = \sigma^2 I$ vs $\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$

Posterior distribution:

$$\mathbf{f}_* | X, \mathbf{y}, x_* \sim \mathcal{N}(\overline{\mathbf{f}_*}, \mathrm{cov}(\mathbf{f}_*)),$$

$$\overline{\mathbf{f}_*} = \mathbf{k}_*^\top (K + \Sigma)^{-1} \mathbf{y},$$

$$\mathrm{cov}(\mathbf{f}_*) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top (K + \Sigma)^{-1} \mathbf{k}_*$$




C. Rasmussen and C. Williams. GP for ML. The MIT Press, 2006.

# Sparse GPR: subset of regressors

✘ Computational cost of a standard GPR: $\mathcal{O}(n^3)$.

**Subset of Regressors** (SR) approach:

- Select a subset of $m$ **representative** points, where $m \ll n$:
  - using farthest-point sampling (FPS), CUR decomposition, etc.
- Use an approximation of the kernel function:

$$k(\mathbf{x}, \mathbf{x}') \approx \mathbf{k}_m(\mathbf{x})^\top K_{mm}^{-1} \mathbf{k}_m(\mathbf{x}')$$

- Compute the predictive mean and variance:

$$\mathbb{E}[f_{\mathrm{SR}}(\mathbf{x}_*)] = \mathbf{k}_m(\mathbf{x}_*)^\top \left(K_{mn}\Sigma^{-1}K_{nm} + K_{mm}\right)^{-1} K_{mn}\Sigma^{-1}\mathbf{y},$$
$$\mathrm{Var}[f_{\mathrm{SR}}(\mathbf{x}_*)] = \mathbf{k}_m(\mathbf{x}_*)^\top (K_{mn}\Sigma^{-1}K_{nm} + K_{mm})^{-1}\mathbf{k}_m(\mathbf{x}_*)$$

✓ Reduced computational cost: $\mathcal{O}(nm^2)$.

# GPR for materials and molecules

**Gaussian Approximation Potentials (GAP)**

- A widely used GPR-based MLIP
- Used for predicting material properties: energies, forces, stresses

Structures $\rightarrow$ Atomic environments $\rightarrow$ descriptors $=: X$
Total energies $\rightarrow$ Local energies $=: Y$

Fit GPR to approximate $f(X) = Y$


Metatrain has a Python implementation of Sparse GAP using Smooth Overlap of Atomic Positions (SOAP) descriptors[1]

[1] Metatrain: Training and evaluating machine learning models for atomistic systems. Metatensor Contributors. https://github.com/metatensor/metatrain, 2025. Accessed: 2025-01-31.

# Kohn-Sham Density Functional Theory (DFT)

Mathematical problem underlying DFT:

$$\min_{\rho} E[\rho]$$

$E[\rho]$ – non-linear DFT energy functional
$\rho(\mathbf{r}) = \sum_{n=1}^{N_{\text{el}}} |\psi_n(\mathbf{r})|^2$ – electron density

Kohn-Sham equations:

$$\left[-\frac{1}{2}\nabla^2 + V_{\text{ext}} + V_{\text{H}} + V_{\text{xc}}\right] \psi_i = \epsilon_i \psi_i$$

$$\epsilon_1 \leq \epsilon_2 \leq \ldots, \quad \langle \psi_n, \psi_m \rangle = \delta_{nm}$$

# Plane-wave discretization

Plane-wave basis set:

$$e_{\mathbf{G}}(\mathbf{r}) = |\Omega|^{-\frac{1}{2}} \exp(i\mathbf{G}\cdot\mathbf{r}), \quad \mathbf{G} \in \mathcal{R}^*.$$

For practical calculations the basis is truncated using an energy cutoff $E_{\text{cut}}$:

$$X_{E_{\text{cut}}} := \text{Span}\left\{ e_{\mathbf{G}}, \quad \mathbf{G} \in \mathcal{R}^*, \quad \frac{1}{2}|\mathbf{G}|^2 \leq E_{\text{cut}} \right\}.$$

# Plane-wave discretization

Plane-wave basis set:

$$e_{\mathbf{G}}(\mathbf{r}) = |\Omega|^{-\frac{1}{2}} \exp(i\mathbf{G}\cdot\mathbf{r}), \quad \mathbf{G} \in \mathcal{R}^*.$$

For practical calculations the basis is truncated using an energy cutoff $E_{\text{cut}}$:

$$X_{E_{\text{cut}}} := \text{Span}\left\{ e_{\mathbf{G}}, \quad \mathbf{G} \in \mathcal{R}^*, \quad \frac{1}{2}|\mathbf{G}|^2 \leq E_{\text{cut}} \right\}.$$

Sources of error in DFT calculations:

- Model error
- **Discretization error**
- Algorithmic error and arithmetic error

✓ Discretization error can be controlled by tuning numerical parameters

✓ Recent methods allow practical estimation of discretization error[1]

[1] Eric Cancès, Geneviève Dusson, Gaspard Kemlin, and Antoine Levitt. Practical error bounds for properties in plane-wave electronic structure calculations. SIAM Journal on Scientific Computing, 44(5):B1312–B1340, 2022.

# Research Objectives

1. Develop an error-informed GPR model.
   - Implement heteroscedastic GPR models and test on simple examples.
   - Integrate heteroscedastic GPR into the Gaussian Approximation Potential (GAP) framework.
   - Optimize computational cost while maintaining predictive accuracy.
2. Incorporate discretization error estimates into the heteroscedastic GAP framework.

# Research Objectives

1. Develop an error-informed GPR model.
   - Implement heteroscedastic GPR models and test on simple examples.
   - Integrate heteroscedastic GPR into the Gaussian Approximation Potential (GAP) framework.
   - Optimize computational cost while maintaining predictive accuracy.
2. Incorporate discretization error estimates into the heteroscedastic GAP framework.
3. Develop a multitask GPR framework for simultaneous prediction of material properties and associated uncertainties.
4. Validate the developed approach on broader datasets.

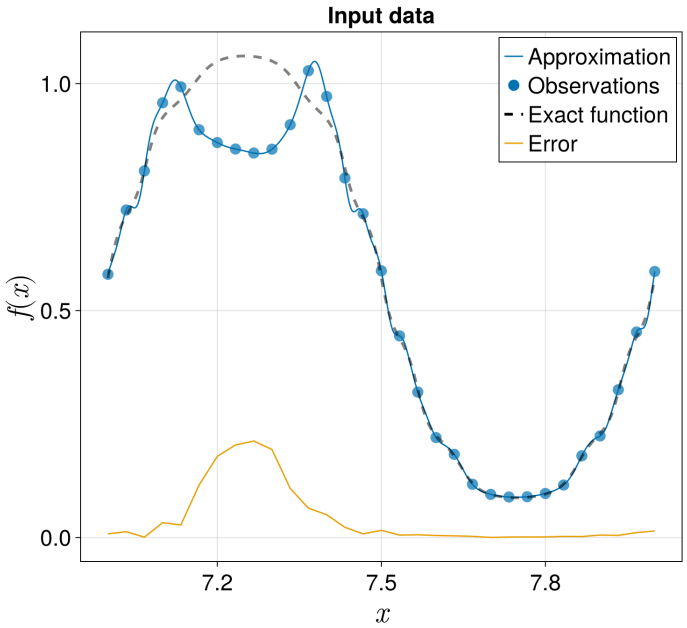
# Results

# Proof of principle: numerical integration

Given a function

$$f(x) = \int_{-3}^{\sin(2\pi x)} \exp\left(-\left(\sin\left(xz\right)\right)^2 - z^2\right) dz,$$

**Goal:** to approximate $f(x)$ with a GP regression using the knowledge about the actual heteroscedastic error inherent to numerical quadrature.



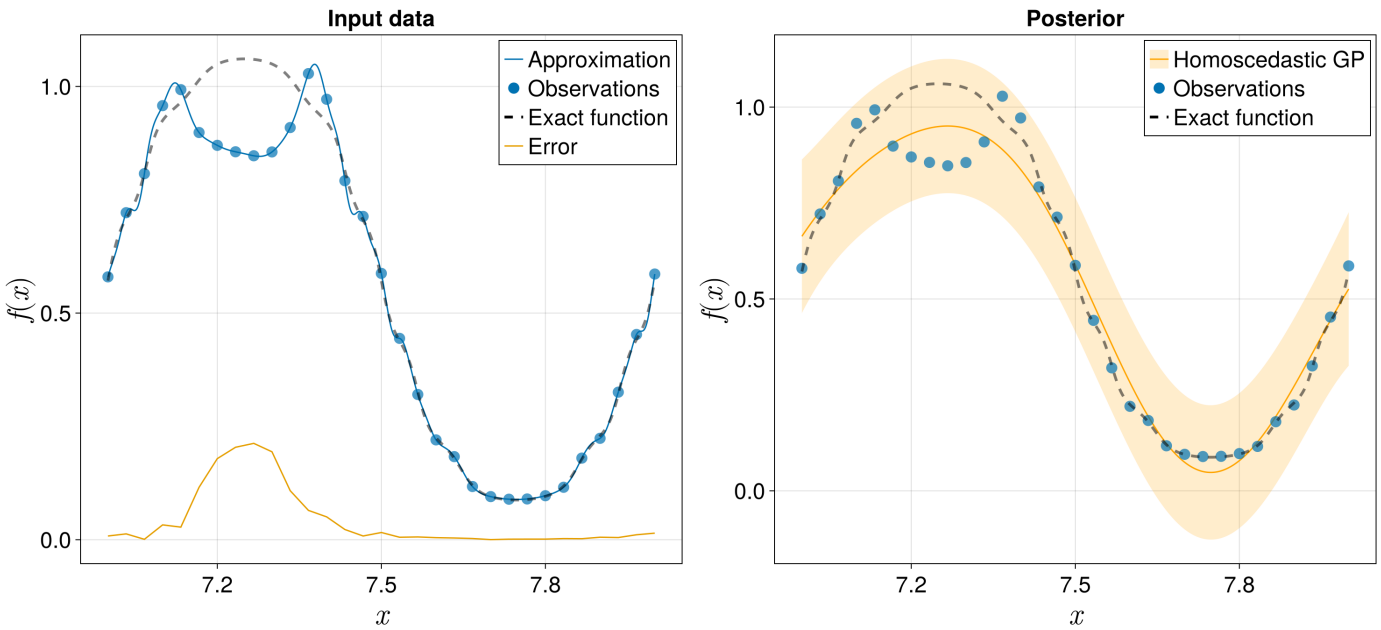

Heteroscedastic GPR model provides **more accurate** approximation

# Proof of principle: numerical integration

Given a function

$$f(x) = \int_{-3}^{\sin(2\pi x)} \exp\left(-\left(\sin\left(xz\right)\right)^2 - z^2\right) dz,$$

**Goal:** to approximate $f(x)$ with a GP regression using the knowledge about the actual heteroscedastic error inherent to numerical quadrature.



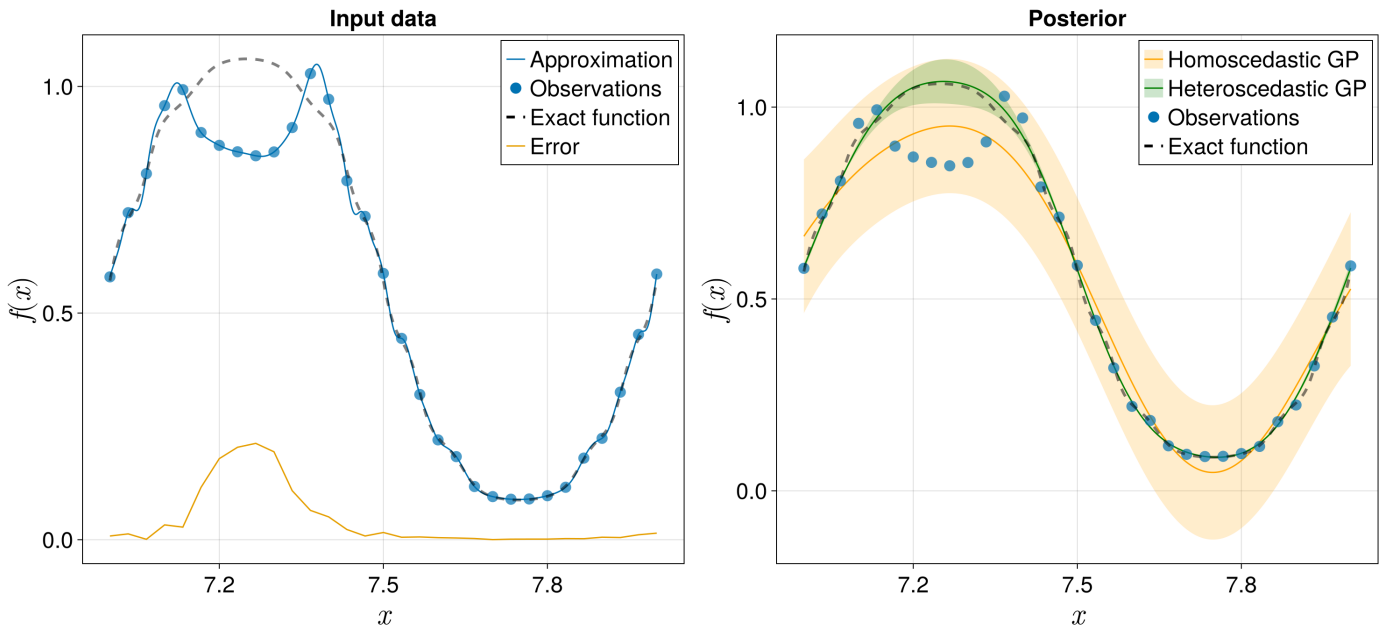

Heteroscedastic GPR model provides **more accurate** approximation

# Proof of principle: numerical integration

Given a function

$$f(x) = \int_{-3}^{\sin(2\pi x)} \exp\left(-\left(\sin\left(xz\right)\right)^2 - z^2\right) dz,$$

**Goal:** to approximate $f(x)$ with a GP regression using the knowledge about the actual heteroscedastic error inherent to numerical quadrature.




Heteroscedastic GPR model provides **more accurate** approximation

# Prediction of material property: equation of state

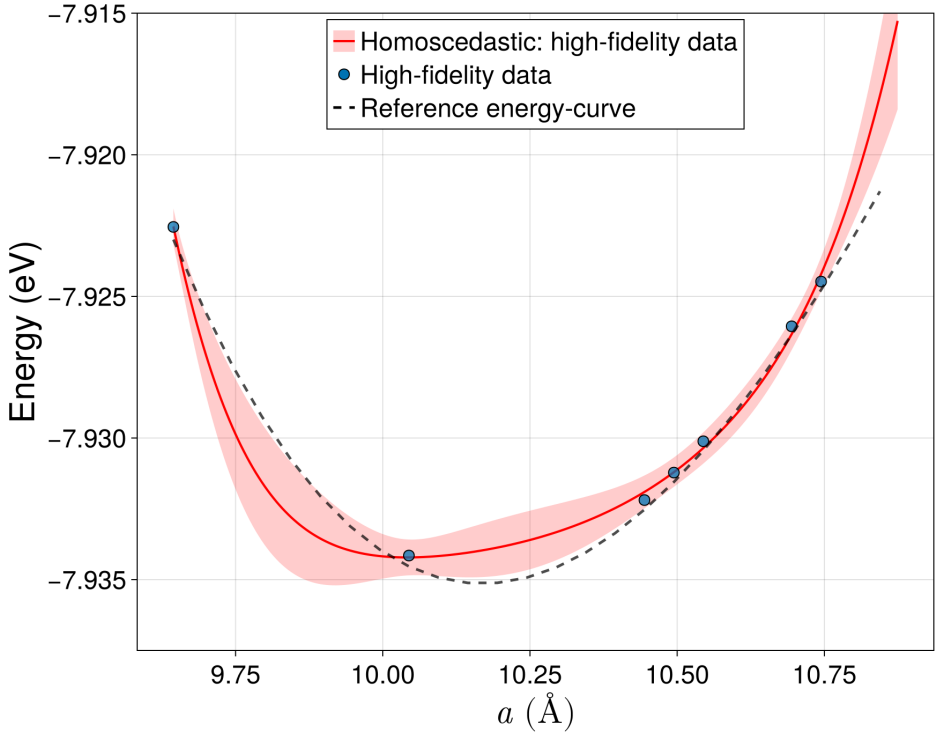
Describes the behavior of a solid under compression/expansion.

**Goal:** to approximate the energy-volume curve with a GP regression for a mixed dataset with varying plane-wave cutoff during data generation.

# Prediction of material property: equation of state

Describes the behavior of a solid under compression/expansion.

**Goal:** to approximate the energy-volume curve with a GP regression for a mixed dataset with varying plane-wave cutoff during data generation.

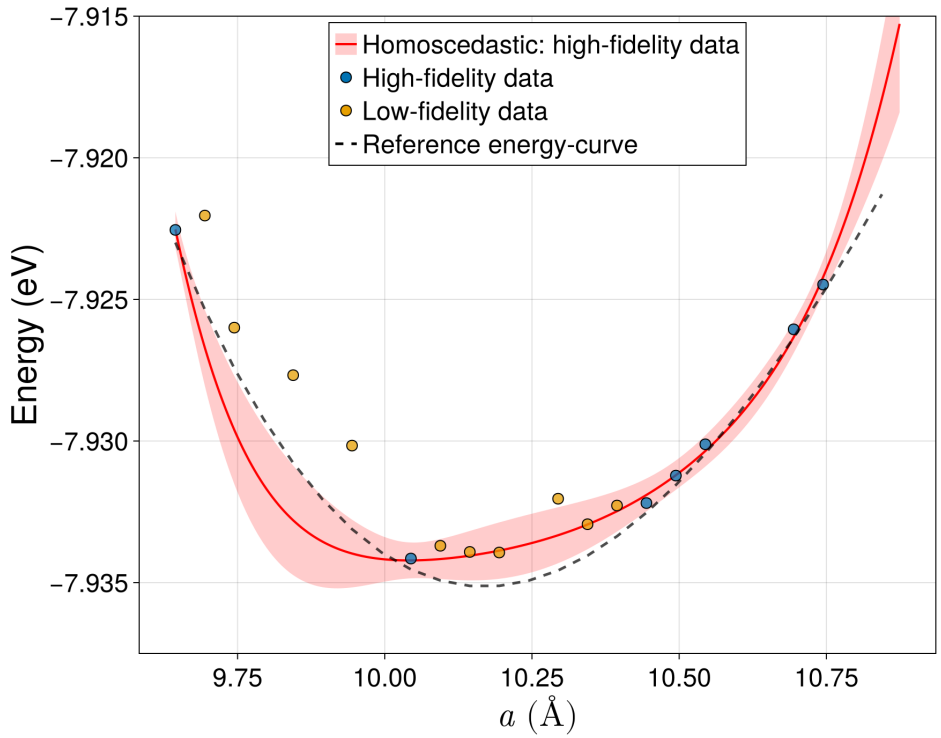



Heteroscedastic GPR model provides **more accurate** approximation

# Prediction of material property: equation of state

Describes the behavior of a solid under compression/expansion.

**Goal:** to approximate the energy-volume curve with a GP regression for a mixed dataset with varying plane-wave cutoff during data generation.

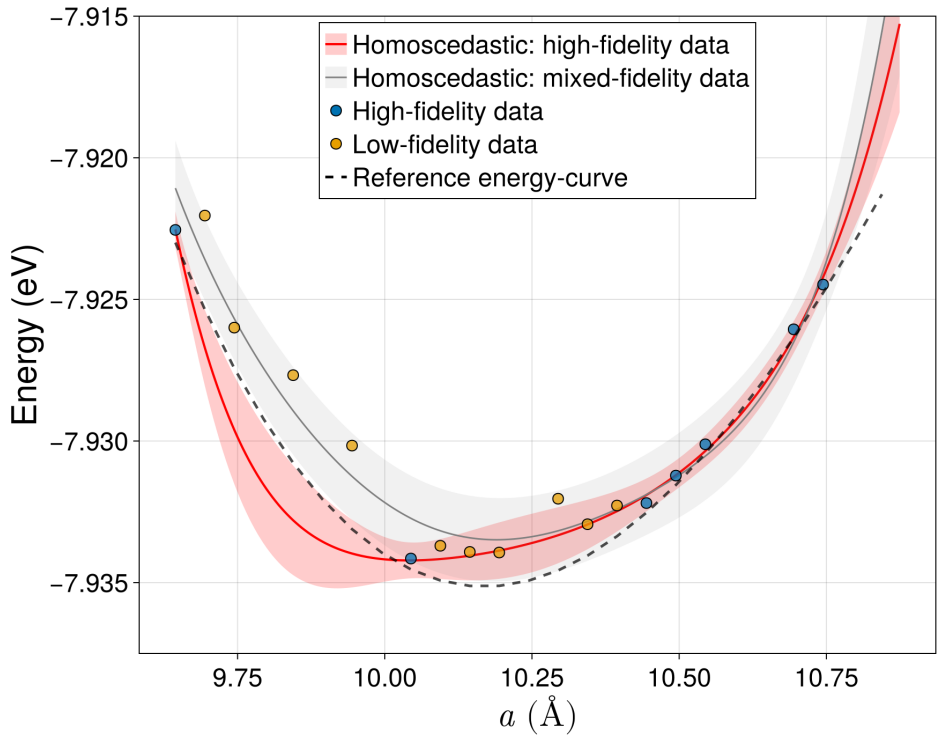



Heteroscedastic GPR model provides **more accurate** approximation

# Prediction of material property: equation of state

Describes the behavior of a solid under compression/expansion.

**Goal:** to approximate the energy-volume curve with a GP regression for a mixed dataset with varying plane-wave cutoff during data generation.

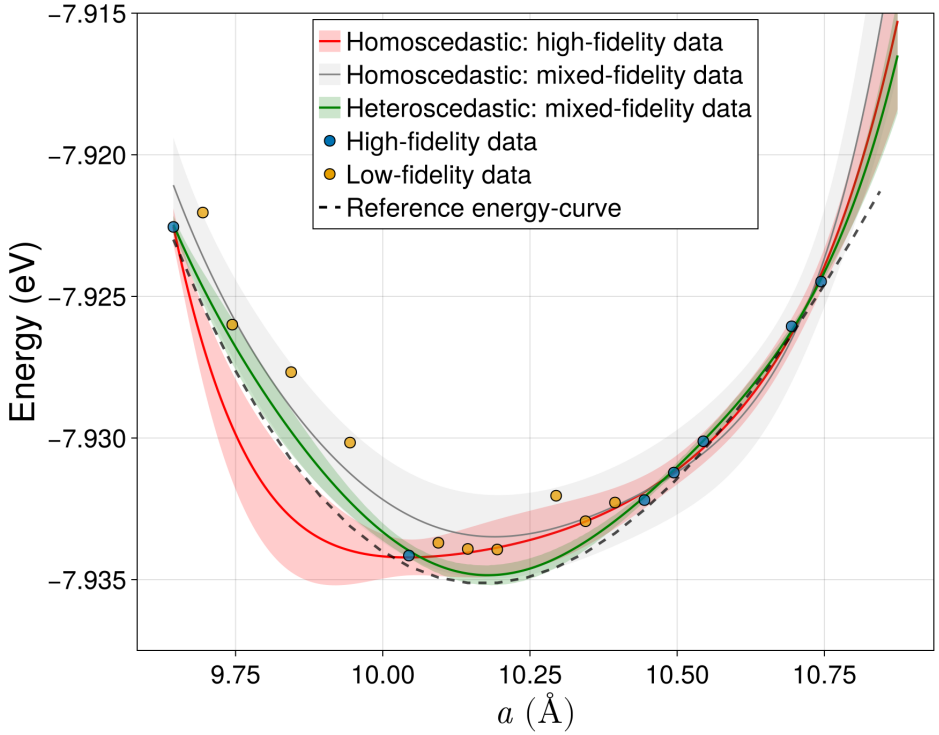



Heteroscedastic GPR model provides **more accurate** approximation

# Prediction of material property: equation of state

Describes the behavior of a solid under compression/expansion.

**Goal:** to approximate the energy-volume curve with a GP regression for a mixed dataset with varying plane-wave cutoff during data generation.

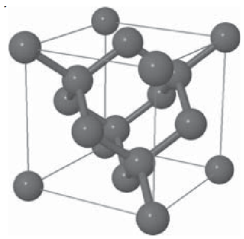
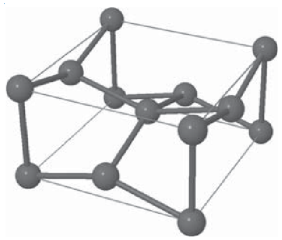
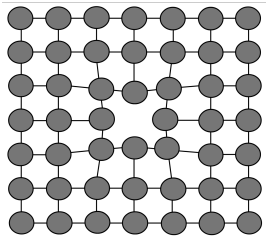



Heteroscedastic GPR model provides **more accurate** approximation

# Large-scale dataset of silicon structures

**Data**: subset of GAP training dataset[1].
**Target properties**: energies and forces.

**Goal:** to approximate energies and forces with a heteroscedastic GPR method incorporated into the GAP model.

Control over the discretization error:

- Plane-wave cutoff energies ($E_{\text{cut}}$): 10 Ha, 18 Ha, and 26 Ha.

[1] Albert P. Bartók, James Kermode, Noam Bernstein, and Gábor Csányi. Machine learning a general-purpose interatomic potential for silicon. Phys. Rev. X, 8:041048, Dec 2018.

# Large-scale dataset of silicon structures

| **Data** | **Diamond** | **$\beta$-Sn** | **Vacancy** | **# of env-s** |
|---|---|---|---|---|
| **High-fidelity** train | 25 | - | - | 50 |
| **Mixed-fidelity** train | 30 | 25 | 5 | 495 |
| Test | 115 | - | 8 | 2162 |

Energy cutoff values ($E_{cut}$):

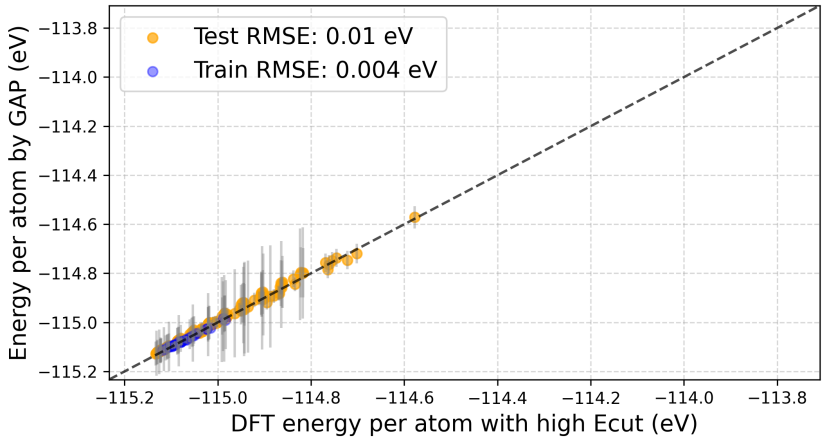
- **High-fidelity** data: 26 Ha
- **Mixed-fidelity** data: 26, 18, 10 Ha (25, 25, 10 structures)

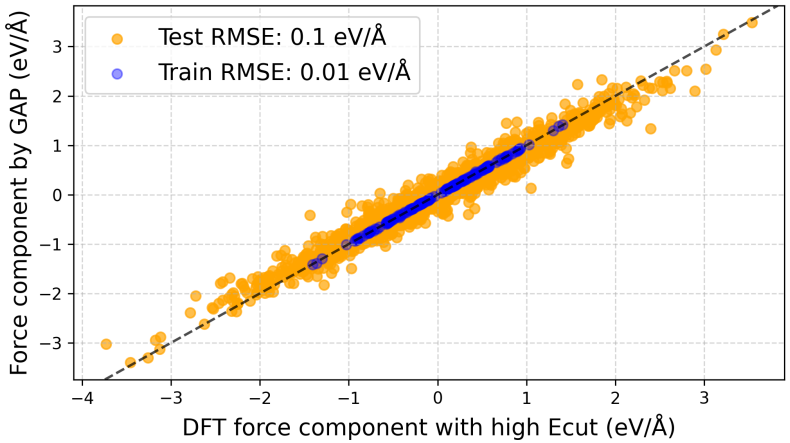
Errors: the difference from reference calculations at 26 Ha

Master student: Victor Bugnion

# Homoscedastic GAP

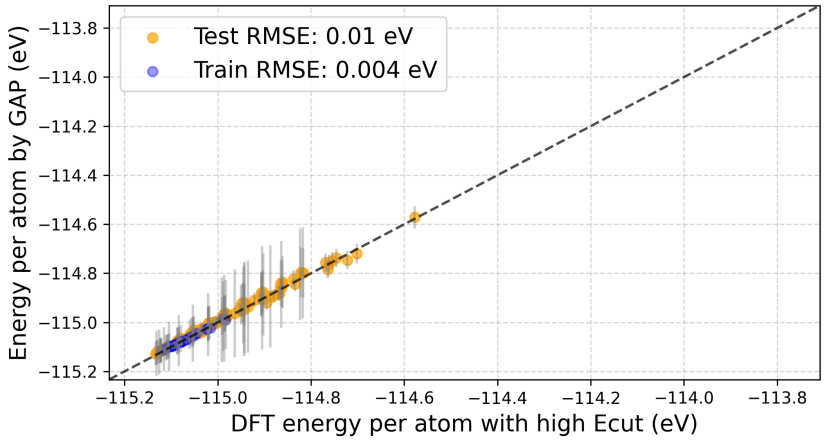
## Trained on high-fidelity data

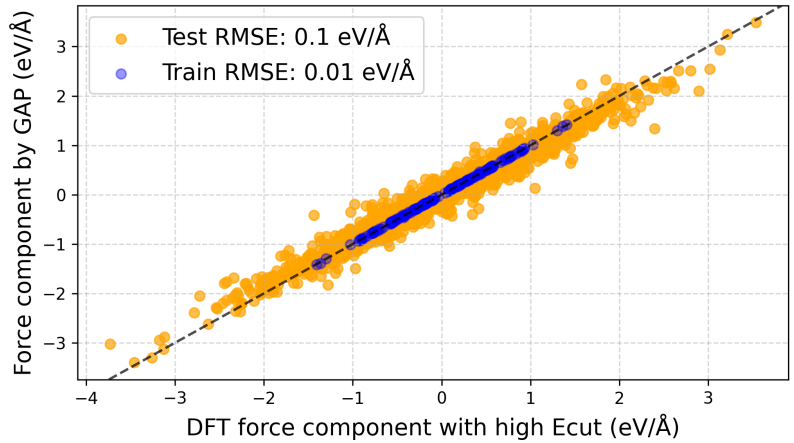


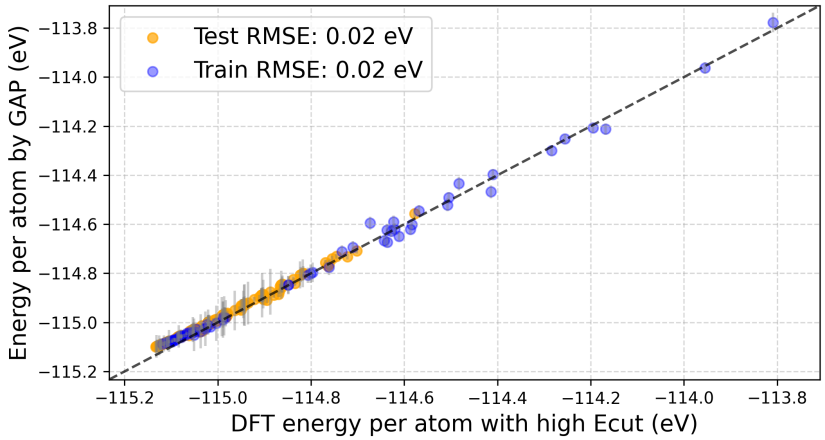
# Homoscedastic GAP

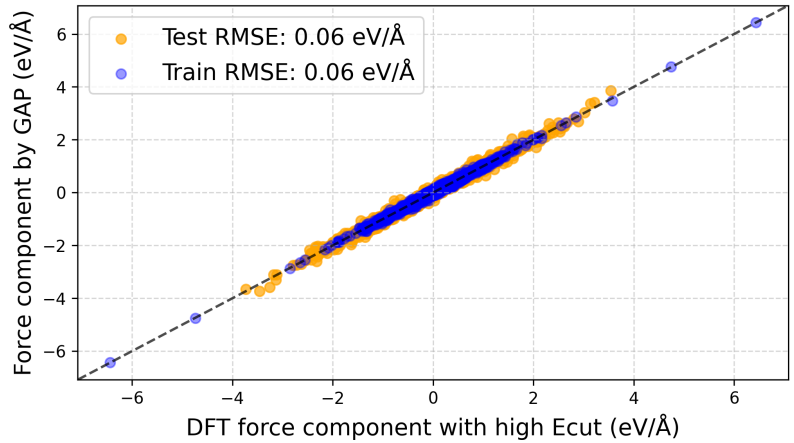
## Trained on high-fidelity data



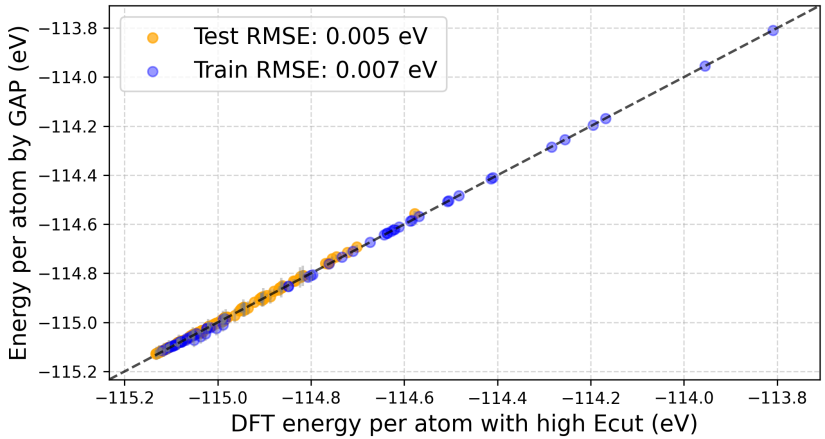

## Trained on mixed-fidelity data

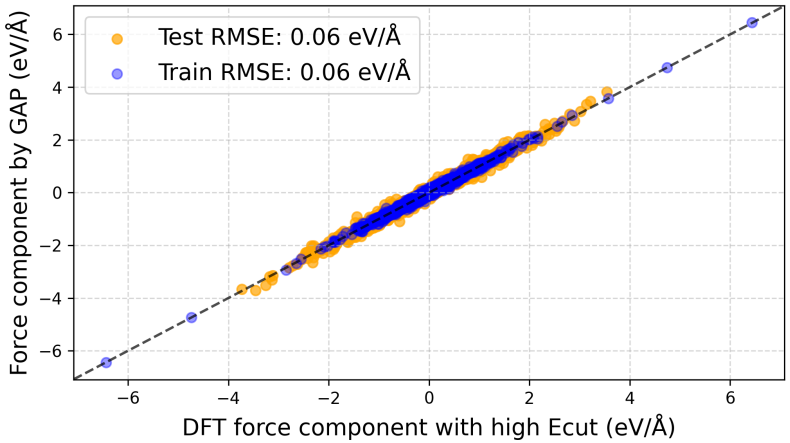


# Heteroscedastic GAP

Trained on mixed-fidelity data with incorporated energy errors




✓ Lower RMSE: 0.02 → **0.005** eV

✓ Improved uncertainty: 6.5% → **0.8**% overconfident predictions

✓ Reduced data generation: using mixed-fidelity dataset

# Future steps: error-informed GPR model

- Evaluate on more complex materials properties (e.g. vacancy formation energy, vacancy migration).
- Include the discretisation errors as part of $\sigma(\mathbf{x_i})$ and test different noise models, for example:

$$\sigma(\mathbf{x_i}) = \alpha + \beta \cdot \delta Q(\mathbf{x_i}), \tag{1}$$

  where $\alpha$ is a bias hyperparameter, $\beta$ is a scaling coefficient, and $\delta Q(\mathbf{x_i})$ represents an estimated discretization error.
- Incorporate **active learning** strategies to improve training efficiency by selectively querying high-uncertainty regions.
- Implement in `scikit-matter`[1] together with Joe Abbott (COSMO).
- Result in publication 1.

Goscinski et al. scikit-matter: A Suite of Generalisable Machine Learning Methods for Chemistry and Materials Science, Open Res. Europe, 3, 81, 2023.

# Future steps: beyond independent noise assumption

- Develop GPR model to simultaneously predict a DFT quantity and its discretization error. In this case, consider one GP for the ground-state properties:

$$Q(\mathbf{x}) \sim \mathcal{GP}(\mu_q(\mathbf{x}), k_q(\mathbf{x}, \mathbf{x}')), \tag{2}$$

and another GP as discretization error:

$$\delta Q(\mathbf{x}) \sim \mathcal{GP}(\mu_\delta(\mathbf{x}), k_\delta(\mathbf{x}, \mathbf{x}')), \tag{3}$$

# Future steps: beyond independent noise assumption

- Consider the **latent variable GPR**, which introduces latent variables that modulate the covariance function, making it suitable for modeling non-stationary and multi-modal processes[1].
- Key idea – to introduce noise in the input of a GP:

$$f(\mathbf{x}) = g(\mathbf{x}, \mathbf{h}) \sim \mathcal{GP}$$

- After discussions with Prof. Carl Henrik Ek (University of Cambridge)
- Result in publication 2.

[1] Bodin et al. Modulated Bayesian Optimization using Latent Gaussian Process Models, NeurIPS. 2020.

# Future steps: error-informed multitask GPR

- Explore refining low-fidelity predictions using a hierarchical model:

$$y^{(\text{high})}(x) = \alpha \cdot y^{(\text{low})}(x) + \eta(x), \tag{4}$$

$$\delta y^{(\text{high})}(x) = \beta \cdot \delta y^{(\text{low})}(x) + \rho(x). \tag{5}$$

- GP prior:

$$\left(y^{(.)}(\mathbf{x}), \delta y^{(.)}(\mathbf{x})\right) \sim \mathcal{GP}(\mu, k), \text{ where } (.) \in \{\text{low}, \text{high}\} \tag{6}$$

- Incorporate **model error** into the learning process.
- Results in potential publication 3.

# Datasets

- Validate developed methods on a **large-scale dataset** of 6 virtual oxide structures and 4 unary systems covering each element of the periodic table[1].
- This dataset is now integrated into the AiiDA workflow manager and is accessible from DFTK.

[1] Bosoni, E., Beal, L., Bercx, M. et al. How to verify the precision of density-functional-theory implementations via reproducible and universal workflows. Nature Reviews Physics. 2023. DOI: https://doi.org/10.1038/s42254-023-00655-3

# Conclusion

- Developed a **heteroscedastic** GPR model capable of handling mixed-fidelity datasets, incorporating error directly into the learning process.

# Conclusion

- Developed a **heteroscedastic** GPR model capable of handling mixed-fidelity datasets, incorporating error directly into the learning process.
- Demonstrated its **effectiveness** on simple examples:
  - numerical integration,
  - equation of state prediction for symmetric silicon structure.

# Conclusion

- Developed a **heteroscedastic** GPR model capable of handling mixed-fidelity datasets, incorporating error directly into the learning process.
- Demonstrated its **effectiveness** on simple examples:
  - numerical integration,
  - equation of state prediction for symmetric silicon structure.
- Extended the methodology, showing its capability on a broader range of silicon structures, confirming that heteroscedastic GPR
  ✓ **improves the prediction power** of MLIPs,
  ✓ **reduces the cost** of data generation.

# Acknowledgment




# Questions?

# Timeline

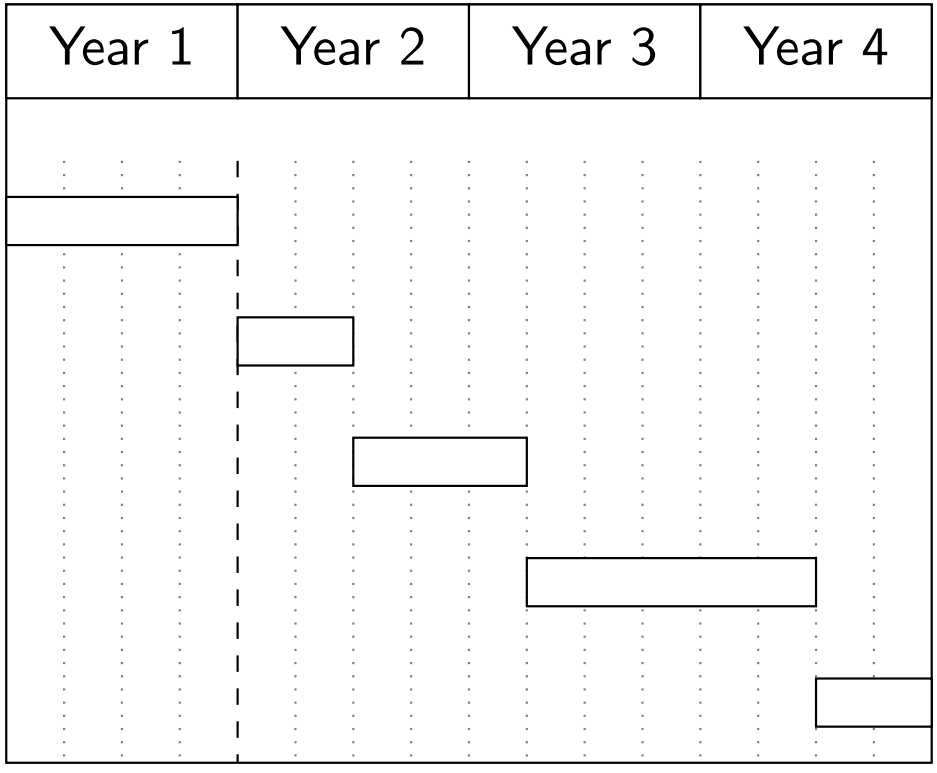
Preliminary Work

Error-Informed GPR Model

Beyond Independent Noise

Uncertainty-Aware Multitask GPR

Thesis Writing



# Large-scale dataset of silicon structures

DFT calculations details:

- Plane-wave cutoff energies ($E_{\text{cut}}$): 10 Ha, 18 Ha, and 26 Ha.
- Maximum k-point spacing: 0.1 inverse Bohrs (around $2\pi \cdot 0.03$ Å$^{-1}$).
- SCF convergence criterion: $\Delta\rho < 10^{-6}$.
- Temperature: $10^{-3}$ K.
- Exchange-correlation functional: PBE.
- Pseudopotential: PseudoDojo for silicon.

Albert P. Bartók, James Kermode, Noam Bernstein, and Gábor Csányi. Machine learning a general-purpose interatomic potential for silicon. Phys. Rev. X, 8:041048, Dec 2018.

# Error distribution

- Energy per atom error:
  - Mean = 0.09 eV
  - Std = 0.23 eV
- Force component error:
  - Mean = **0.002** eV/Å
  - Std = **0.0012** eV/Å

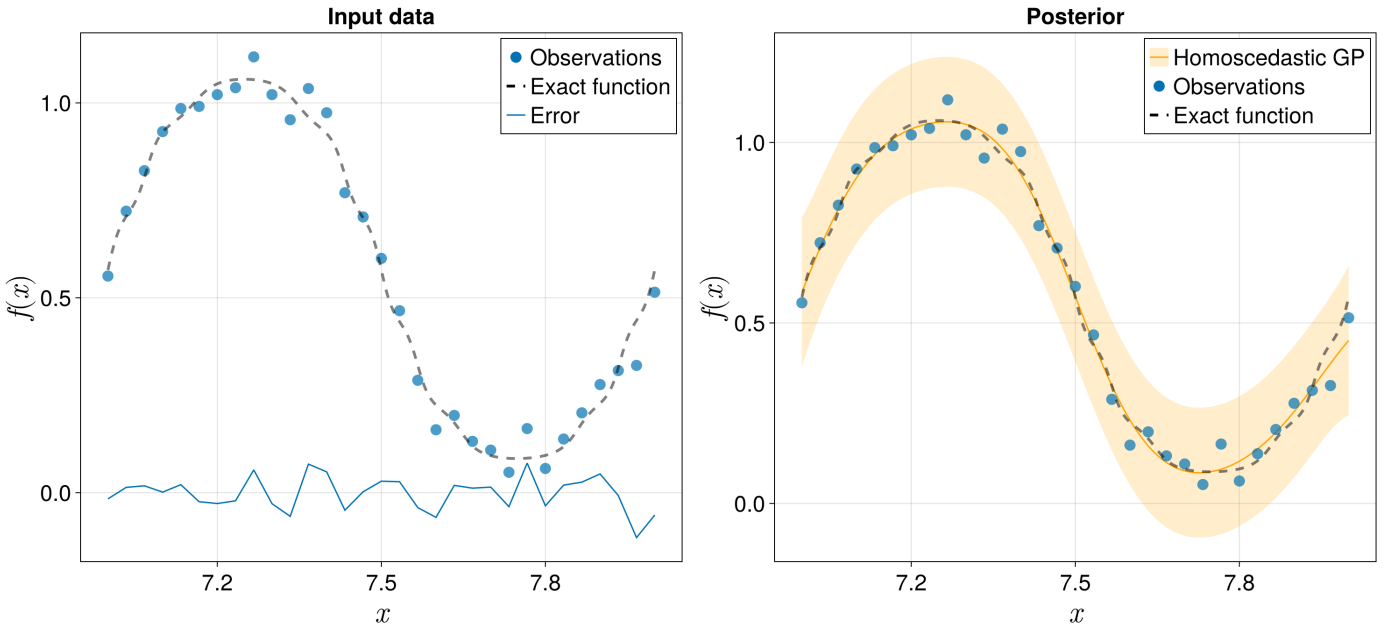
# Proof of principle: numerical integration

Given a function

$$f(x) = \int_{-3}^{\sin(2\pi x)} \exp\left(-\left(\sin\left(xz\right)\right)^2 - z^2\right) dz,$$

**Goal:** to approximate $f(x)$ with a GP regression using the knowledge about the actual heteroscedastic error inherent to numerical quadrature.

**Ideal world**: small Gaussian noise



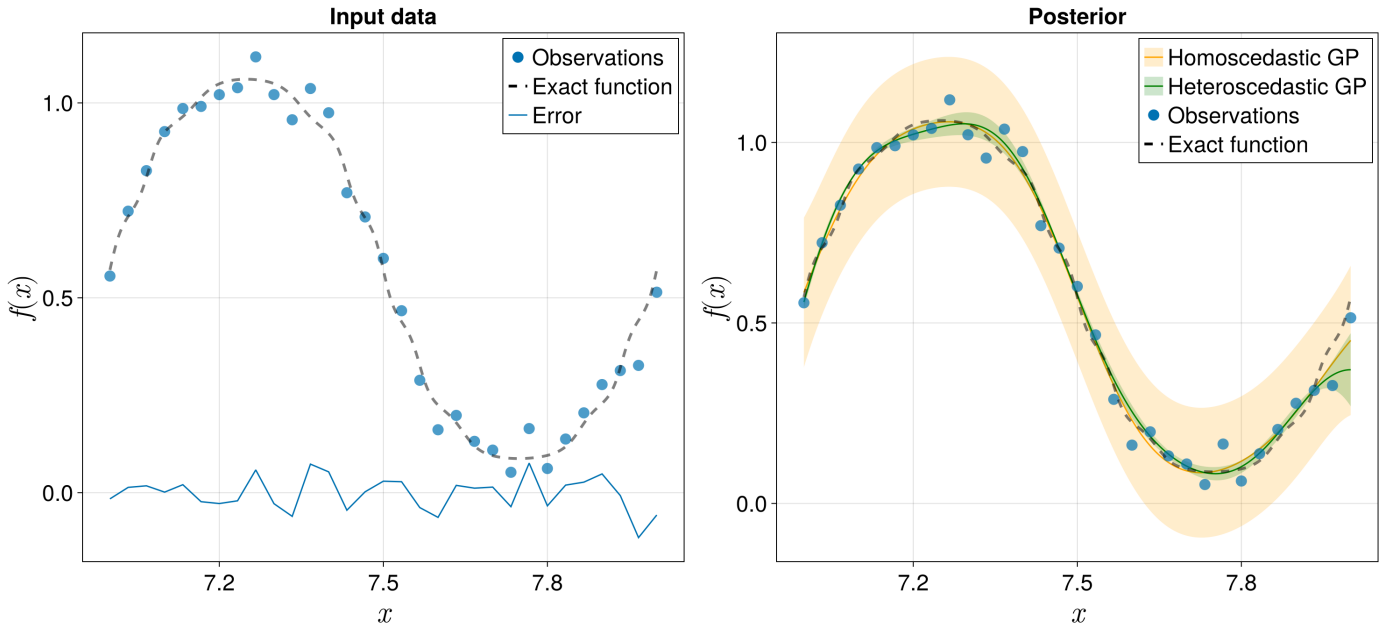
# Proof of principle: numerical integration

Given a function

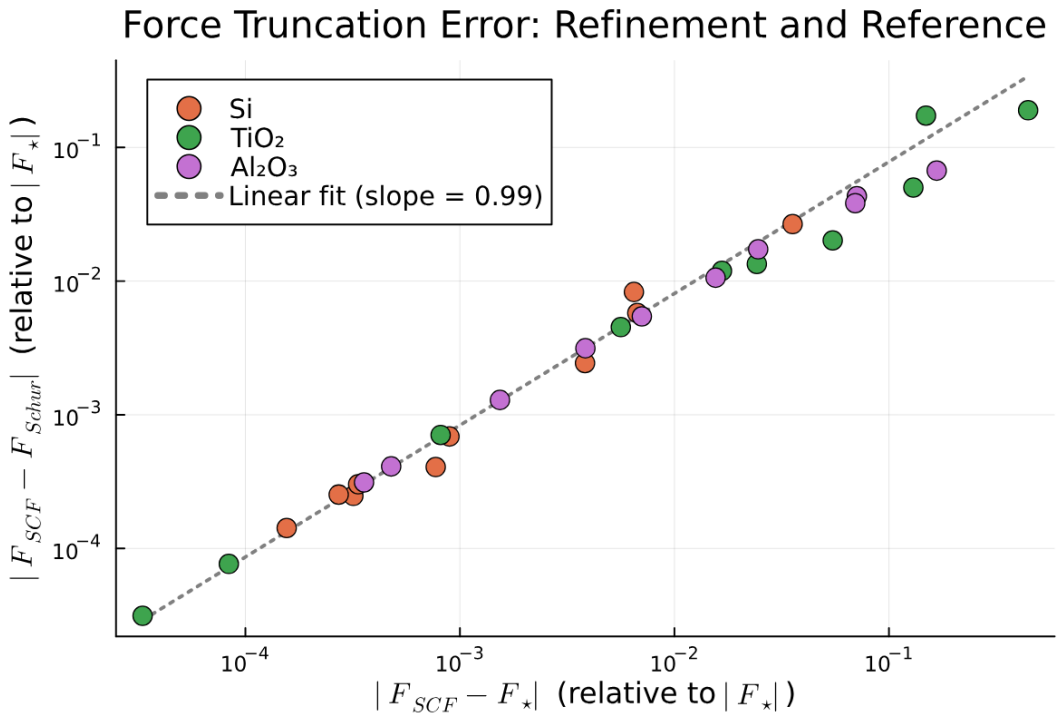
$$f(x) = \int_{-3}^{\sin(2\pi x)} \exp\left(-\left(\sin\left(xz\right)\right)^2 - z^2\right) dz,$$

**Goal:** to approximate $f(x)$ with a GP regression using the knowledge about the actual heteroscedastic error inherent to numerical quadrature.

**Ideal world**: small Gaussian noise



# Numerical error estimates




Estimation of basis set error in forces

PhD student: Bruno Ploumhans