Prediction of Si/Ge Melting Point Curves using Gaussian Process Regression

Anusha Chowdhury

Received September 05, 2024
Accepted November 20, 2024
Electronic access November 30, 2024

Identifying the melting curves of semiconductors such as Germanium (Ge) and Silicon (Si) is important in not only academic or research fields but also in technology and its related fields. Standard experimental techniques, as accurate as they may be, are time-consuming and expensive. This paper suggests a computational strategy based on Ab Initio Molecular Dynamics (AIMD) simulations as well as Gaussian Process (GP) Regression for predicting these melting curves. By employing automated thermodynamic integration, we show good precision of the predictions can be obtained along with their associated uncertainties, which are vital for gauging the credibility of the outcomes. The paper also proves that GP Regression is capable of modeling nonlinear systems in the field of material science at a significantly lesser cost as compared to carrying out a lot of experimentation. For verification of the used methodology, the simulated melting curves are compared with the existing experimental data and the quality of the results defined by the confidence intervals is analyzed. My research thus highlights the potential of integrating AI with computational simulations to advance material design, reducing the dependence on costly experimental approaches and paving the way for more efficient methods in predicting material properties.

Introduction

The melting point of materials, especially semiconductors like Germanium (Ge) and Silicon (Si), is crucial for both basic research and practical uses. However, deciding on such curves with the help of experimental means can be rather time-consuming and expensive. Traditional procedures are less practical due to the requirement for exact environmental control, and the difficulties in collecting reliable measurements, particularly under extreme conditions. In addition, they are also not as computationally efficient. A current method based on Ab Initio Molecular Dynamics (AIMD)[1] with Gaussian Process Regression (GPR) gives the ability to estimate the melting curve in the wide temperature-pressure interval by computing a single melting point. The general complexity of computing the melting point is an order of magnitude higher than the complexity of derivative dataset computations[2]. The GPR based model provides not only the melting curve but can address the uncertainty estimation problem and estimate properties that rely on computations of free energy derivatives. Additionally, this new approach provides convergence with respect to the number of atoms in the AIMD simulations. These simulations rely on quantum mechanics to model atomic behavior and to forecast material characteristics when subjected to certain conditions.

Recent developments in using Artificial Intelligence (AI) in materials science have also provided promising directions for automating processes and enhancing efficiency. Out of the developed AI techniques that can be applied in this regard, Gaussian Processes (GP) Regression[3] has been noted to yield very desirable results. GP is particularly helpful for the analysis of complex or dynamic and non-linear systems since it is good at predicting and modeling these systems. The GP is vastly used in the field of machine learning interatomic potential construction[4]. GP-based atomic descriptors show state-of-the-art performance in fitting potential energy landscapes of molecules and complex materials. This proves the ability of GPs to be excellent approximators of nonlinear functionals. This feature is extremely useful when the data is expensive to obtain or limited, such as in the case of melting curves derived from AIMD simulations. Thus, dealing with GP allows achieving accurate predictions, as well as estimation of the amount of uncertainty related to these predictions, which is vitally important for scientific purposes and can guide further experimentation and/or simulation studies.

GP has been used successfully in effectively predicting material properties such as phase diagrams[5] transport properties[6], material stress or strain[7], etc. in some prior work. We extend this field of work by presenting in this paper a computational method that synthesizes Gaussian Process Regression to generate melting curves of Ge and Si using AIMD, along with automated Thermodynamic Integration (TI) methods. In the following study, the goal will be to show that GP does not only allow for forecasting the melting points of these semiconductors but also allows for estimating the confidence of these forecastings. The study hypotheses that GP, when combined with AIMD simulations, will yield

precise melting curves accompanied by measurable uncertainty estimates, which will significantly decrease the demand for experimental validations and provide a more effective avenue for materials development. While the GPR is a commonly used machine learning technique, its application to materials properties prediction is limited to interatomic descriptors construction[3] as stated earlier. One of the major advantages of the current model is that the kernel of the Gaussian process regression is physically informed. It accounts for the temperature and volume dependence of the free energy of the solid and liquid phases.

Thus, even though other forms of Machine Learning techniques such Neural Network (NN) and Support Vector Machines (SVM) exist, our choice of GPR is motivated by the fact that GPR is a nonparametric model that allows easy inclusion of the physical behavior, while Neural Networks depend on a large number of parameters and SVM provides less flexibility. Both of them require an extensive training set for the model fitting compared to GPR.

The paper is divided into four main parts. Section 1 outlines the data set along with visualization charts and preprocessing steps essential for accurate modeling. Additionally, the section introduces Gaussian Process Regression and how it can be applied to model material properties using Thermodynamic Integration (TI) methods, which are critical for calculating the melting points of Germanium (Ge) and Silicon (Si). In Section 2, the results are presented, showcasing the predicted melting points of Ge and Si across various pressures. These predictions are compared with experimental data to evaluate the model's accuracy. In section 3, we discuss the significance of the results, understand reasons for some inaccuracies and explore possibilities for some future improvements. Section 4 concludes the paper by summarizing the work and emphasizing the successful integration of GPR and TI in predicting melting points and the significance of this method in material science.

MATERIALS AND METHODS

The data collected in this research work was obtained from a computational AIMD simulation and experimental findings concerning the thermal properties of Ge and Si. Specifically, the data comprised Temperature and Volume dependent Energy and Pressure values at various atomic configurations near the melting point obtained from AIMD simulations. This information is necessary to determine the appropriate changes in Ge and Si characteristics as it transfers from a solid state to a liquid state.

The data includes:

- Energy and Pressure Data (sol_fcc.dat & liq.dat): These files include values of Free energy gradient w.r.t Temperature (dF/dT) and Free energy gradient w.r.t Volume (dF/dV) of Ge and Si in its solid and liquid states respectively at various Volumes (V), Temperature (T) and number of particles (N) that were captured during simulation procedure. The dF/dT and dF/dV presented are obtained in the form of potential energy (E) and pressure (P).
- Reference Data (E0.dat): This file contains reference values of Temperature E0 and Pressure (P0 = -dE0/dV) at 0 Kelvin Temperature and this applies to only the solid state of Ge and Si.

The simulations were performed using density functional theory (DFT) calculations which are considered to be very accurate in predicting materials properties at the atomic level. These simulations captured the complex interactions between atoms in the crystalline structure of Ge and SI by generating data at several temperatures near the predicted melting point. In addition, the AIMD simulations were generated with machine learning potential drained on reference data obtained with density functional theory using VASP software. The trained potential was used to perform classical molecular dynamics simulation with the LAMMPS software package to sample energy and pressures from statistically independent cuts of the trajectory.

Data Post-Processing

Before applying the GPR model, the raw data underwent several critical post-processing steps to ensure accuracy in the predictions:

Temperature Conversion:

Temperature in the raw data collected is in Kelvin (K). It is converted to electron volts (eV) using the Boltzmann constant. This conversion is necessary as the GP model was parameterized in eV in order to maintain consistency with physical energy units. Additionally in thermodynamics, the ensemble properties of number of particles (N), volume (V) and temperature (T) are naturally obtained as derivatives of the βF, where F is the Helmholtz free energy and β=1/kT. Hence, it is reasonable to keep βF unitless, which is achieved by having T in the units of eV.

$$T_{\mathrm{eV}} = T_{\mathrm{K}} \times k_B$$

where ($k_B = 8.617 \times 10^{-5}$, eV/K

Normalization and Error Handling

Energy and Pressure Normalization:

This was done on the raw data obtained from the experiments, by subtracting the reference energy at 0 K (E0) and dividing the difference by the square of the temperature to avoid

divergence as the temperature approaches zero. Likewise, the pressure values were normalized using the above formula: by subtracting the reference pressure (P0) from it, and applying a temperature-dependent scaling factor.

Error Estimation:

The energy and pressure standard deviations were computed for every temperature point, giving an estimate of measurement of the uncertainty or error. To make the errors found here consistent with the scaled pressure and energy numbers, they were also normalized. The normalized values were computed as follows:

$$Y_{\text{Energy}} = \frac{E - E_0}{T^2} - \frac{3}{2T}$$

$$Y_{\text{Pressure}} = \frac{P - P_0}{T} + \frac{1}{V}$$

The -3/2T term in the energy equation corresponds to the limit of free energy derivative of the solid phase with respect to temperature in the limit of 0 temperature. The 1/V term in the pressure equation is the ideal gas limit for the liquid phase. These limits are subtracted to ensure normal distribution of the target variable.

Feature Construction:

The data was organized into feature vectors comprising temperature, volumes, and number of particles. All of these features were taken as inputs in the GPR model. Derivative Computation: The derivatives of the GP kernel with respect to volume and temperature were calculated symbolically. This is needed as the output vector contains Free Energy derivative w.r.t Temperature and Volume.

Kernel Function Optimization:

Using the post-processed data, the squared exponential kernel and its derivatives were optimized. The function values and gradients were taken into consideration when adjusting the hyperparameters, in order to get the best fit for the normalized energy and pressure data. The GPR is a nonparametric model which means hyperparameters of the model can be found exactly in the training procedure by maximizing log likelihood function. Due to the second term in the log-likelihood function, GPR is well known to be a model that does not suffer from overfitting. The optimization procedure in each case was performed until convergence of the target function up to parameter 1e-8 between the iterations using the L-BFGS-B algorithm.

Data Visualization

Figure 1 to 4 provide a good visualization of the Ge and Si data set contained in their respective files, named sol_fcc.dat and E0.dat, which contain their Energy and Pressure in solid state at different temperatures and volume.

Gaussian Process (GP) in General

Gaussian Process (GP) regression is a Bayesian non-parametric method of data modeling. It offers a way of estimating the behavior of unknown functions using the observed data in terms of probability. As opposed to the previous methods of parametric modeling where a certain shape of a function is assumed, GPs provide a probability distribution over possible functions taking into consideration all possible behaviors and their respective uncertainties. A GP is fully specified by its mean function, m(x), and covariance function, k(x,x'):

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

Mean function m(x):

Represents the expected value of the function at each point x. It is often assumed to be as zero as possible for choice simplicity especially if there is no given background information concerning the function involved.

Covariance function (Kernel) k(x,x'):

Explains how the function is related at two distinct points/venues x and x'. The kernel function along with its hyperparameters influences the smoothness, periodicity, and other characteristics of the functions that the model tries to fit.

The most frequently used kernel for covariance function is called the Squared Exponential Kernel which is shown below. The Squared Exponential Kernel function gives a good generalization for the normally distributed variable. According to thermodynamics, for a system with an infinite number of particles the free energy of the system is a Gaussian distributed variable. To ensure the target distribution has a mean value close to zero, we subtracted the zero temperature limit for solid and the ideal limit for liquid. Other kernels such as Matern and periodic kernels are designed for specific tasks while the Gaussian kernel is general and can be adjusted to include physical inside in the functional form.

$$k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2l^2}\right)$$

Here, σ^2 is the variance of the output, and l is the length scale, controlling how quickly the correlation between points decreases with distance.

Multivariate Gaussian Process Regression

In many practical applications, the quantity of interest is not a single scalar function but rather multiple correlated outputs, such as different physical properties of a material that depend on shared underlying factors. Multivariate Gaussian Process (MGP)

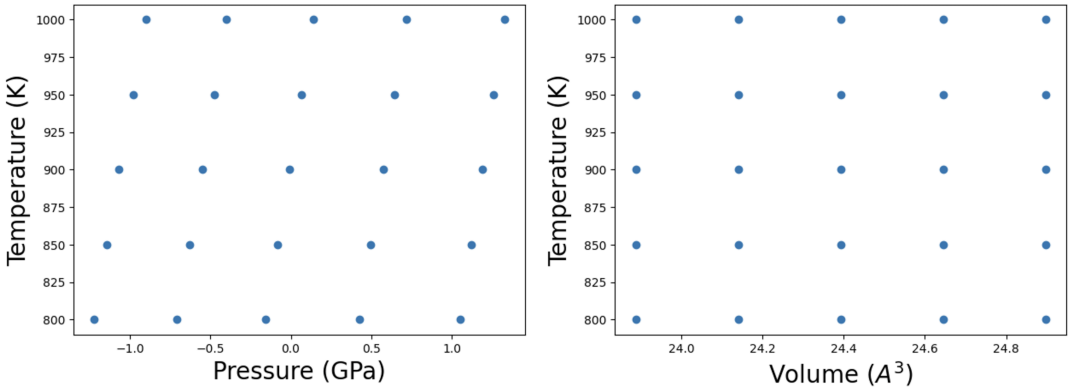

Figure 1: Visualization of Ge data set: sol_fcc.dat

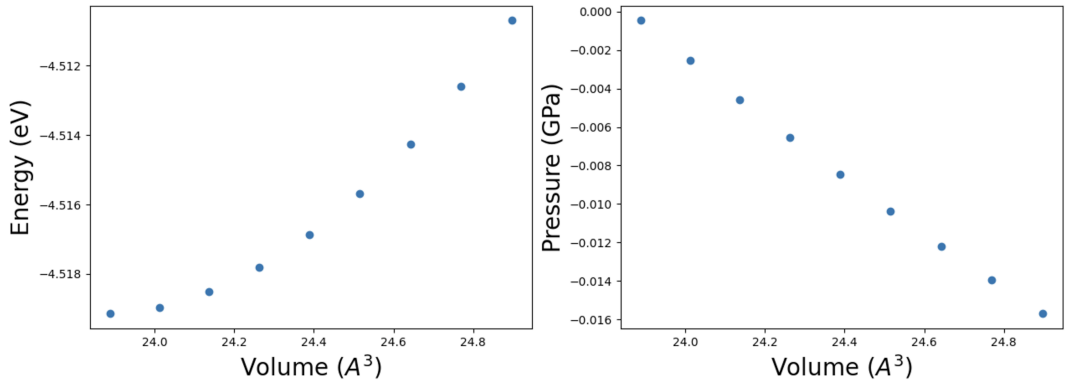

Figure 2: Visualization of Ge data set: E0.dat

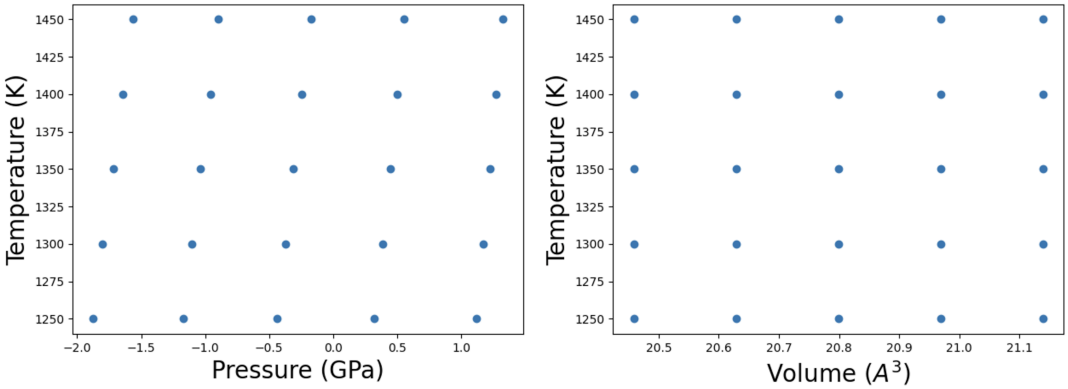

Figure 3: Visualization of Si data set: sol_fcc.dat

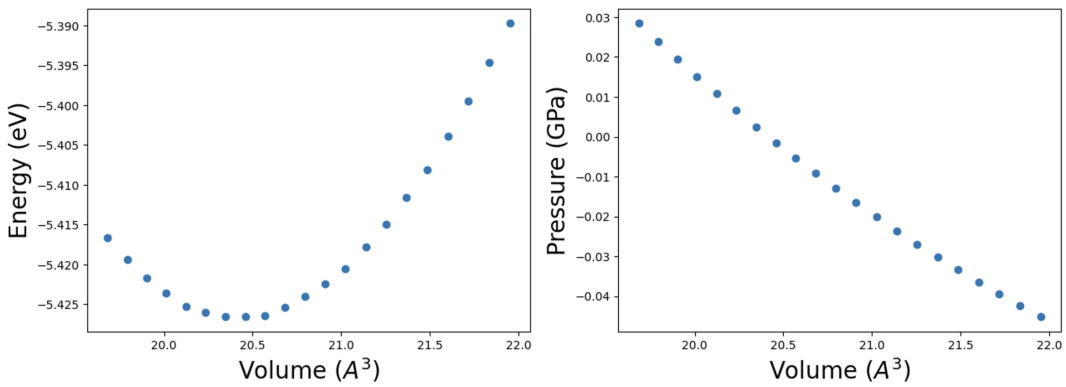

Figure 4: Visualization of Si data set: E0.dat

regression extends the GP framework to model multiple outputs simultaneously, taking into account the correlations between them[8]. Below are the multivariate Kernel functions we used for solid and liquid state of the two materials.

Here x1 = V1, N1, T1 and x2 = V2, N2, T2 are two input data points having

For solid:

$$k_{\text{GPR\_E0}}(x_1,x_2) = s^2 \exp\left(-\frac{(1/V_1 - 1/V_2)^2}{2l_V^2}\right)$$

Here, $x_1 = \{V_1\}$ and $x_2 = \{V_2\}$ are two input data points having V_1 and V_2 volumes, respectively.

$$k_{\text{GPR\_(TVN-Solid)}}(x_1,x_2) = s_0^2 + s^2 \exp\left(-\frac{(T_1 - T_2)^2}{2l_T^2}\right)$$

$$\cdot \exp\left(-\frac{(1/V_1 - 1/V_2)^2}{2l_V^2}\right) \cdot \exp\left(-\frac{(1/N_1 - 1/N_2)^2}{2l_N^2}\right)$$

Here, $x_1 = \{T_1, V_1, N_1\}$ and $x_2 = \{T_2, V_2, N_2\}$ are two input data points having T_1 and T_2 temperatures, V_1 and V_2 volumes, and N_1 and N_2 number of particles, respectively.

For liquid:

$$k_{\text{GPR\_(TVN-liquid)}}(x_1,x_2) = s_0^2 + \frac{s_1^2}{T_1 T_2} + s^2 \exp\left(-\frac{(T_1 - T_2)^2}{2l_T^2}\right)$$

$$\cdot \exp\left(-\frac{(1/V_1 - 1/V_2)^2}{2l_V^2}\right) \cdot \exp\left(-\frac{(1/N_1 - 1/N_2)^2}{2l_N^2}\right)$$

Here x1 = T1, V1, N1 and x2 = T2, V2. N2 are two input data points having T1 and T2 temperatures, V1 and V2 volumes and N1 and N2 number of particles respectively. Using these kernels, we trained the GPR models of Ge and Si with the data set we described earlier. Table 1 lists the hyperparameters that were obtained after training.

Multivariate Gaussian Process Regression with Derivative

In many applications, the data set contains partial derivatives of the output w.r.t the some of its input dimension rather than the actual output themself. For example, the dataset we used for Gen and Si and obtained from finite length molecular dynamic simulation contains Free Energy gradient w.r.t Temperature (i.e. proportional to internal energy E) and Free Energy gradient w.r.t Volume (i.e. proportional to pressure P). Having outputs in the form of gradients can greatly improve the data fit of the Gaussian Process model. When we are dealing with derivative outputs, the covariance matrix K includes not only the covariance between function values but also the cross-covariances involving derivatives[9].

Modeling Accuracy:

We used the above described method of multivariate Gaussian Process Regression with derivatives to model and fit the input data set. Table 2 summarizes model accuracies in terms of relative errors and R2 scores. The metrics were obtained through 5-fold cross-validation with splitting of the dataset into train and test parts with a ratio of 4:1 (shuffling of the data was performed before the splitting). Since energy and pressure values in the dataset have different scales we divide the MAE and RMSE metrics by the mean of the corresponding derivative (MAEE and MAEP). A visual way to understand the modeling accuracy is to plot the input dataset and overlay them with model computed values. We used this technique to plot the below figures:

- Figure 5 overlays the Ge solid Pressure vs Temperature and Energy vs Temperature points in the dataset with model computed values.

Material	State	Model	Hyperparameters		
			l_T	l_V	l_N
Ge	Solid	GPR_E0	NA	0.03299926	NA
	Solid	GPR_TVN_sol	0.58380731	-0.01976855	14.75425612
	Liquid	GPR_TVN_liq	0.58380731	-0.01976855	14.75425612
Si	Solid	GPR_E0	NA	0.02710855	NA
	Solid	GPR_TVN_sol	0.53496202	-0.01420259	7.3965012
	Liquid	GPR_TVN_liq	0.53496202	-0.01420259	7.3965012

Table 1: Hyperparameters of the trained GPR models

- Figure 6 overlays the Ge liquid Pressure vs Temperature and Energy vs Temperature points in the dataset with model computed values.
- Figure 7 overlays the Si solid Pressure vs Temperature and Energy vs Temperature points in the dataset with model computed values.
- Figure 8 overlays the Si liquid Pressure vs Temperature and Energy vs Temperature points in the dataset with model computed values.

As can be seen from Figure 5 through 8 and Table 2, the models have very good accuracy in fitting the input data set.

GP Applied to Thermodynamic Properties Estimation

Gaussian Process regression provides a very powerful method for the estimation of thermodynamic properties. In materials science, thermodynamics properties like heat capacity, thermal expansion coefficients, and melting points etc. are affected by atomic interactions that are difficult to describe with traditional means. GP excels in such scenarios with several advantages:

Modeling Non-linear Behavior:

Thermodynamic properties may depend on temperature, pressure and composition nonlinearly [10]. The modeling of these non-linearities is easier in GPs because GPs do not require the precise specification of a certain functional form. The data used to estimate the thermodynamic properties contain a certain level of uncertainty due to the randomness in experiments or in MD simulations. GPs give a probabilistic measure of this uncertainty, and provide standard deviations alongside point predictions.

Estimating the Melting Point Using GP:

At the melting point, the below three equations need to be satisfied which provide relations between the energies for solid and liquid state at a common Pressure and Temperature point. With the help of the created GPR models, these equations were precisely represented and by using python's non-linear equation solver, we solved these equations to get the Temperature and Pressure points at which these equations satisfy. These solution points represent the melting point of Ge and Si. In our case we are interested in finding melting point T as a function of pressure P. In terms of free energy F of solid (subscript sol) and liquid (subscript liq) phases system of equations has a form

$$\frac{\partial F_{\mathrm{sol}}(T,V)}{\partial V} = -P$$

$$\frac{\partial F_{\mathrm{liq}}(T,V)}{\partial V} = -P$$

$$F_{\mathrm{sol}}(T,V) - F_{\mathrm{liq}}(T,V) = P\left(V_{\mathrm{liq}} - V_{\mathrm{sol}}\right)$$

In the equations Vsol and Vliq are volumes of solid and liquid phase at the melting point

RESULTS

Figure 9 and 10 shows the computed melting points of Ge and Si at various Pressures computed by the work we presented. This is overlaid with real melting points data. In these charts, the Tref is the melting temperature added to the dataset. For the red curve, it corresponds to the computational melting point and for the green curve, the melting point is taken from the existing experimental data. As can be seen, our model based result compares very well with real experiment based data. The P(T) dependence is well reproduced by the GPR while the DFT tends to underestimate the melting temperature compared to the experimental data. The difference seen is due to the inaccuracies in AIMD based data.

The density functional theory used for machine learning potential fitting is the approximate Ab Initio model. Its predictions can diverge from the real system behavior. DFT tends to underestimate the melting temperature by 100K-200K. They can be mitigated by the inclusion of the experimental melting temperature in the dataset as shown in the paper or

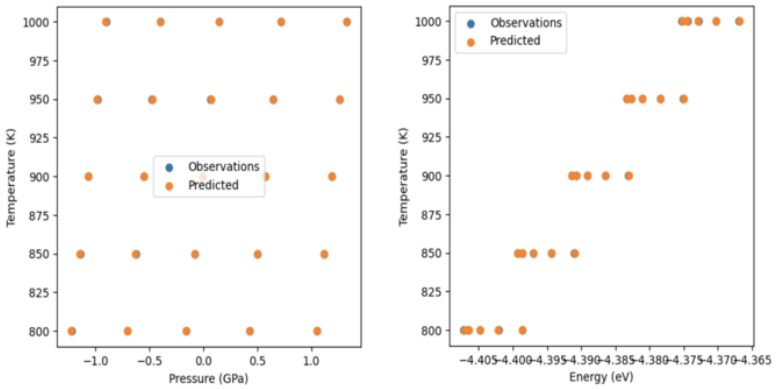

Figure 5: Ge Solid Model predicted vs Observed values

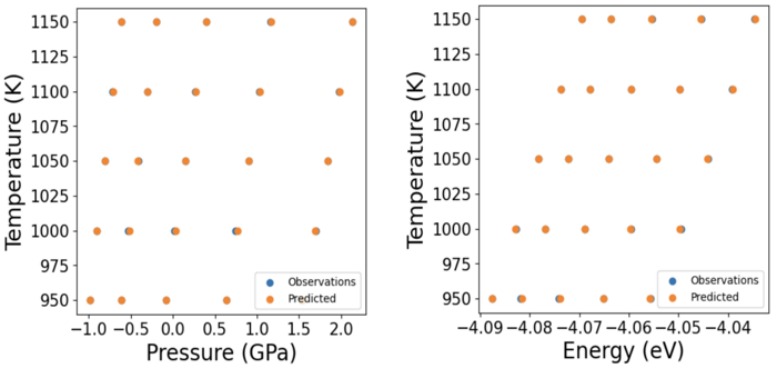

Figure 6: Ge Liquid Model predicted vs Observed values

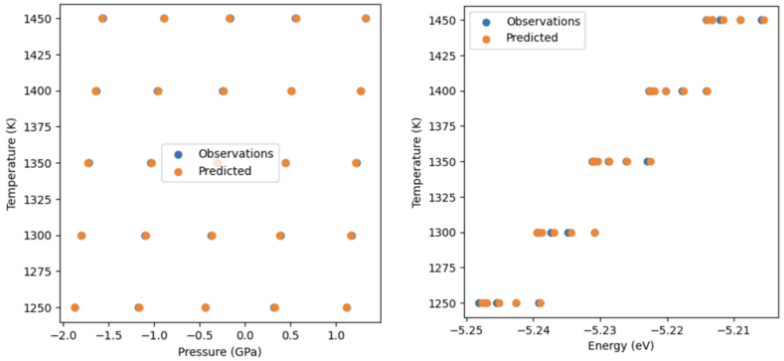

Figure 7: Si Solid Model predicted vs Observed values

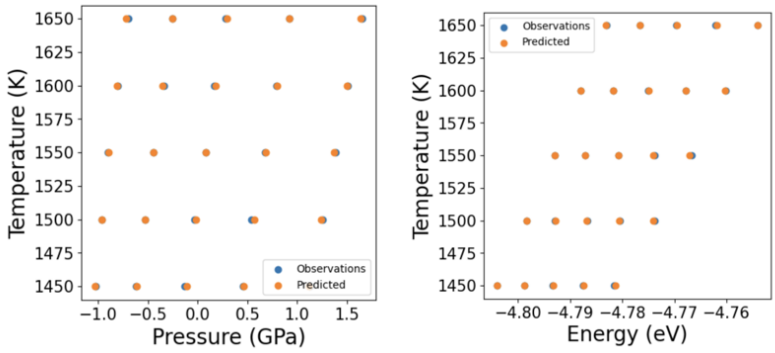

Figure 8: Si Liquid Model predicted vs Observed values

Material	State	Model	Metrics					
			MAE_E (%)	MAE_P (%)	RMSE_E (%)	RMSE_P (%)	R2_E	R2_P
Ge	Solid	GPR_E0	0.0003(1)	0.24(3)	0.0003(1)	0.24(3)	0.999994(2)	0.9996(3)
Ge	Solid	GPR_TVN_sol	0.89(4)	0.56(2)	1.0(1)	0.57(2)	0.998(1)	0.999(4)
Ge	Liquid	GPR_TVN_liq	0.006(3)	0.9(3)	0.007(2)	1.10(5)	0.999999(2)	0.9999(7)
Si	Solid	GPR_E0	0.002(1)	0.16(4)	0.0014(6)	0.1(1)	0.99994(3)	0.99998(4)
Si	Solid	GPR_TVN_sol	0.9(5)	0.5(3)	1.1(5)	0.6(3)	0.997(1)	0.998(3)
Si	Liquid	GPR_TVN_liq	0.006(4)	0.9(1)	0.007(3)	1.1(1)	0.99997(1)	0.9999(3)

Table 2: Models performance metrics.

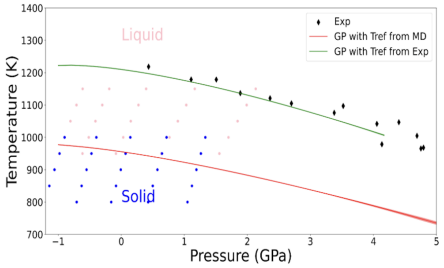

Figure 9: Computed vs experimental Ge melting point plot

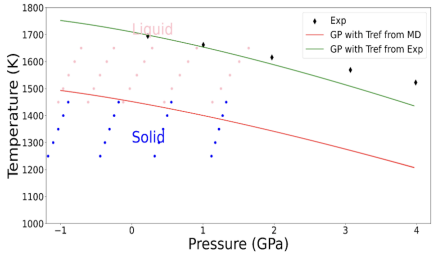

Figure 10: Computed vs experimental Si melting point plot

by using a more accurate Ab Initio model for the interatomic potential training.

The blue and pink points are computational data from MD simulation for solid and liquid phases, correspondingly. The overlap is possible since the melting point is the point on the phase diagram where solid and liquid phases coexist. Without a solid liquid interface solid can be found to be stable above the melting temperature in the MD simulation.

DISCUSSION

These results show that Gaussian Process Regression (GPR) using Molecular Dynamics (AIMD) simulations generated data, is an effective method for predicting the melting points of Germanium (Ge) and Silicon (Si) across various pressures reducing the reliance on costly and time-intensive experimental procedures. The predicted melting points of Ge and Si are very close to the experimental melting curves, thus highlighting the capability of GPR to model complex, non-linear material behaviors.

However, the results also show some discrepancies between the predicted and experimental data. This discrepancy mainly arises from the inherent limitations in the AIMD simulations, which rely on density functional theory (DFT) and introduce noise (or errors) in the dataset that propagate through the GPR model.

Additionally, while the squared exponential kernel was effective, exploring more complex kernels or hybrid modeling approaches could further improve predictive accuracy. Overall, this research highlights the potential of integrating GPR with computational simulations for efficient and accurate materials property prediction, paving the way for advancements in materials science with reduced experimental dependency.

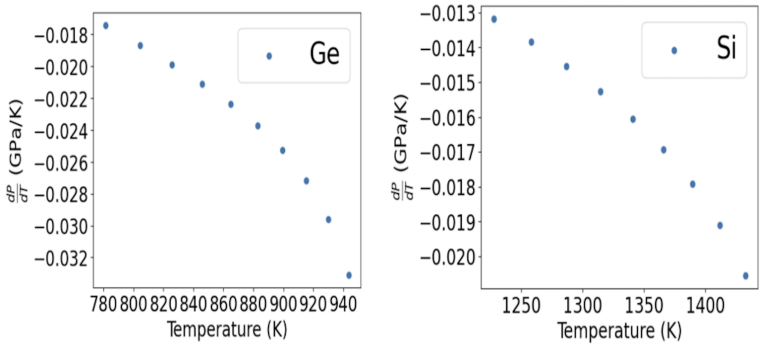

Figure 11: Slope temperature dependence for Si and Ge.

CONCLUSION

In the paper I have applied Gaussian Process regression for the prediction of melting curves of Ge and Si. The technique demonstrated good accuracy in predicting the melting curves of materials in a wide range of pressures based on simulation data. The slope and relative shift of the melting curves agrees with experimental data highlighting the effectiveness of this technique. The discrepancy with respect to experiment is attributed to the drawbacks of the underlying approach of data generation which is based on density function theory. Overall, the present paper shows success of the chosen approach for the estimation of the melting curves of materials.

ACKNOWLEDGMENTS

I express gratitude to Vladimir Ladygin and Lumiere Education Program for their valuable resources, insightful discussions and guidance related to the topic.

References

1 G. Kresse and J. Hafner, *Physical Review B*, 1993, **47**, 558–561.

2 S. Yoo, X. C. Zeng and J. R. Morris, *Journal of Chemical Physics*, 2004, **120**, 1654–1656.

3 Z. Ghahramani and C. E. Rasmussen, *Gaussian processes for machine learning*, MIT Press, 2006.

4 A. Bartók-Pártay, *The Gaussian Approximation Potential: an interatomic potential derived from first principles quantum mechanics*, Springer Science Business Media, 2010.

5 V. Ladygin, I. Beniya, E. Makarov and A. Shapeev, *Physical Review B*, 2021, **104**, 104102.

6 L. Zeni and C. P. Lowe, *Journal of Chemical Physics*, 2020, **152**, 064102.

7 J. Chen, Y. Wu and V. Sundararaghavan, *Computer Methods in Applied Mechanics and Engineering*, 2019, **355**, 412–431.

8 M. Álvarez and N. D. Lawrence, *Journal of Machine Learning Research*, 2011, **12**, 1459–1500.

9 M. Solak, R. Murray-Smith, W. E. Leithead, D. J. Leith and C. E. Rasmussen, Advances in Neural Information Processing Systems 15 (NIPS 2003), 2003, pp. 1057–1064.

10 K. A. Dill and S. Bromberg, *Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology*, Garland Science, 2003.