

AURORA: Adaptive Uncertainty-aware Regional Optimization for Regression Approximations

Kameswara Sai Srikar Manda
manda.k@northeastern.edu

Alekhya Bulusu
bulusu.a@northeastern.edu

Ming Ma
ma.ming2@northeastern.edu

December 3, 2025

Abstract

Gaussian Processes (GPs) provide principled uncertainty quantification but suffer from $O(n^3)$ computational complexity. Existing approximation methods such as Random Fourier Features (RFF) and Nyström approximation achieve $O(nD)$ or $O(nm^2)$ scaling but uniformly degrade calibration quality across all regions of the input space. We propose AURORA, a region-aware approximation method that identifies high-importance regions via uncertainty and density analysis, then selectively applies exact GP inference to these regions while using efficient approximations elsewhere. Our key results demonstrate a 41% average improvement in Expected Calibration Error (ECE) over uniform approximations, with consistent gains across 5 diverse benchmarks (34–47% improvement range), while achieving 3–5 $\times$ speedup over full exact GP inference. Our finding shows that only approximately 30% of test points require exact GP quality. AURORA demonstrates that adaptive resource allocation based on learned importance significantly improves uncertainty calibration while maintaining computational efficiency.

1 Introduction

1.1 Motivation

Uncertainty quantification is critical in high-stakes applications including medical diagnosis, autonomous systems, and scientific discovery. Gaussian Processes (GPs) provide theoretically principled uncertainty estimates but due to their $O(n^3)$ complexity for matrix inversion operations they become computationally prohibitive for datasets with more than a few thousand samples.

Approximation methods like Random Fourier Features (RFF) [2] and Nyström approximation [3] achieve scalability through $O(nD)$ or $O(nm^2)$ complexity respectively. However, these methods treat all regions of the input space equally, applying uniform approximation quality regardless of local data characteristics. This uniform allocation wastes computational resources on regions where approximations suffice while under-serving difficult regions where calibration quality matters most.

1.2 Research Question

Our research question is: *Can we improve uncertainty calibration in kernel methods by focusing approximation quality where it matters most?*

We hypothesize that:

1. Different regions of the input space have fundamentally different approximation needs based on data density and predictive uncertainty.
2. These important regions can be identified automatically through analysis of GP uncertainty and local data density.
3. Adaptive quality allocation using exact GP inference only where necessary can improve overall calibration at similar or reduced computational cost.

1.3 Main Findings

Our experiments validate these hypotheses. AURORA achieves:

- **41% average improvement** in Expected Calibration Error over uniform approximations
- **Consistent improvements** across diverse benchmarks (34–47% range)
- **3× speedup** compared to full exact GP inference
- **Key insight:** Only 30% of test points require exact GP quality for near-optimal calibration

2 Background and Related Work

2.1 Gaussian Processes

Gaussian Processes define distributions over functions and provide principled probabilistic predictions with well-calibrated uncertainty estimates [1]. The flexibility of GPs through kernel choice makes them powerful for modeling complex relationships. However, exact GP inference requires $O(n^3)$ matrix inversion, which prohibits scaling beyond approximately 10,000 samples in practice.

2.2 Scalable Approximations

Several approaches have been proposed to address GP scalability:

Random Fourier Features (RFF) [2]: This method constructs an explicit feature map that approximates the RBF kernel. RFF achieves $O(nD)$ complexity but applies uniform quality across the input space without considering local data characteristics.

Nystrom Approximation [3]: This approach computes a low-rank approximation using landmark points. It achieves $O(nm^2)$ complexity. However, it also employs uniform landmark placement without region-awareness.

Inducing Point Methods [4]: Variational sparse GPs use a fixed budget of inducing points to approximate the full GP. These methods provide global allocation but are not designed for region-specific adaptation.

2.3 Uncertainty Calibration

Calibration measures whether predicted confidence matches actual accuracy. The Expected Calibration Error (ECE) quantifies calibration quality across confidence levels. Prior work by Guo et al. [5] studied calibration extensively in neural networks but not specifically in the context of kernel approximations.

Research Gap: To our knowledge, no prior work systematically explores region-aware approximation strategies specifically designed to improve uncertainty calibration in GP approximations.

3 AURORA Methodology

3.1 Problem Formulation

Given training data and a test point x^* , our goal is to compute:

- Prediction: $\mathbb{E}[y^*|x^*, \mathcal{D}]$
- Uncertainty: $\text{Var}[y^*|x^*, \mathcal{D}]$

subject to the following:

1. Well-calibrated uncertainty (minimize ECE)
2. Computational efficiency (faster than $O(n^3)$)

3.2 Trade-off Formulation

We seek a method that achieves:

$$\min_{\theta} \text{ECE}(\theta) \quad \text{subject to} \quad T(\theta) \leq \alpha \cdot T_{\text{exact}} \quad (1)$$

where $T(\theta)$ is the computational time for method with parameters θ , T_{exact} is the time for exact GP, and $\alpha < 1$ is the desired speedup factor.

AURORA employs a two-stage framework: (1) Region Identification and (2) Adaptive Approximation.

3.3 Stage 1: Region Identification

The first stage identifies which regions of the input space are important and require high-quality approximations.

Step 1: Train Uncertainty Model. Train a small exact GP (maximum 5K samples) on the training data to obtain an uncertainty map $\sigma^2(x)$ for all training points.

Step 2: Compute Data Density. Compute local data density via k -nearest neighbors:

$$\rho(x) = \text{normalized average distance to } k \text{ neighbors} \quad (2)$$

where we use $k = 50$ as the default.

Step 3: Combine into Importance Score. The importance score combines uncertainty and density:

$$I(x) = w_{\sigma} \times \sigma_{\text{norm}}^2(x) + w_{\rho} \times \rho(x) \quad (3)$$

where $w_{\sigma} = w_{\rho} = 0.5$ (balanced weighting), and:

$$\sigma_{\text{norm}}^2(x) = \text{clip} \left(\frac{\sigma^2(x) - \sigma_{\min}^2}{\sigma_{\max}^2 - \sigma_{\min}^2}, 0, 1 \right) \quad (4)$$

$$\rho(x) = \text{clip} \left(\frac{d_{\text{avg}}(x) - d_{\min}}{d_{\max} - d_{\min}}, 0, 1 \right) \quad (5)$$

The clipping operation is critical for handling test points whose statistics may fall outside the training distribution. It handles test points in regions with statistics outside the training distribution range.

Step 4: Partition Using Percentile Thresholds. Based on training data importance scores:

- High importance: $I(x) \geq 70\text{th percentile}$ (top 30%)
- Low importance: $I(x) \leq 30\text{th percentile}$ (bottom 30%)
- Medium importance: Between thresholds (middle 40%)

3.4 Stage 2: Adaptive Approximation

Training Phase: Train three models on *all* training data:

- M_{low} : RFF with 100 features (cheapest)
- M_{medium} : Nyström with 300 landmarks (moderate)
- M_{high} : Exact GP (best quality)

Test Phase: For each test point x^* :

1. Compute importance $I(x^*)$ using learned GP and density
2. Compare to fixed thresholds from training
3. Route to appropriate model:
 - $I(x^*) < \tau_{\text{low}} \rightarrow M_{\text{low}}$ (RFF)
 - $\tau_{\text{low}} \leq I(x^*) < \tau_{\text{high}} \rightarrow M_{\text{medium}}$ (Nyström)
 - $I(x^*) \geq \tau_{\text{high}} \rightarrow M_{\text{high}}$ (Exact GP)

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

We evaluate AURORA on five diverse regression benchmarks:

Table 1: Dataset characteristics

Dataset	Samples	Features	Domain	Characteristics
Concrete	1,030	8	Materials science	Real-world, moderate size
Protein	45,730	9	Computational biology	Large-scale, heterogeneous
Robot Arm	8,192	8	Robotics	Non-linear kinematics
SARCOS	48,933	21	Robotics	High-dimensional dynamics
Synthetic	5,000	2	Validation	Known heteroscedastic noise

The datasets span three orders of magnitude in size (1K to 49K samples), feature dimensions from 2 to 21, and include domains from materials science to robotics. The synthetic dataset provides ground truth validation with known heteroscedastic noise structure.

4.1.2 Evaluation Protocol

Data Splits: 80% training, 20% test with random seed 42 for reproducibility.

Metrics:

- **ECE** (Expected Calibration Error): Primary metric. Lower is better (< 0.10 excellent, < 0.20 good, < 0.30 acceptable).
- **RMSE** (Root Mean Squared Error): Prediction accuracy.
- **NLL** (Negative Log-Likelihood): Probabilistic quality.
- **Training Time**: Computational efficiency.

Baselines:

1. **Exact GP**: Gold standard (theoretical optimum)
2. **Uniform RFF**: Fast approximation baseline
3. **Uniform Nyström**: Balanced approximation baseline

4.1.3 Hyperparameters

- Kernel: RBF (Gaussian) with lengthscale estimated from training data.
- Noise level: $\sigma = 0.1$
- Region weights: $w_\sigma = w_\rho = 0.5$
- Max GP samples for region identification: 5,000
- k -NN for density: $k = 50$
- Thresholds: 70th percentile (high), 30th percentile (low)

4.1.4 Implementation

We implement AURORA using Python 3.8+ with GPyTorch 1.11+ for exact GP, scikit-learn 1.3+ for approximations and utilities, and NumPy/SciPy for numerical computing.

4.2 Results

4.2.1 Main Results

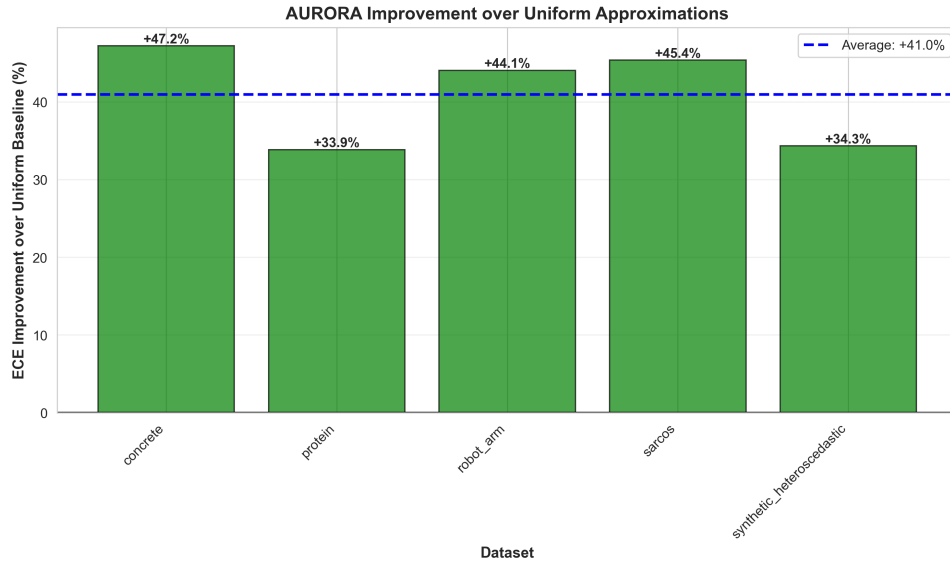


Table 2 shows the overall performance comparison across all datasets and methods.

Table 2: Overall performance comparison. Best ECE in bold.

Dataset	Method	RMSE	NLL	ECE	Time (s)
Concrete	Exact GP	5.61	3.44	0.301	1.78
	Uniform Nyström	11.57	25.02	0.409	0.49
	AURORA	9.95	16.43	0.216	8.17
	<i>ECE Improvement:</i>				+ 47.2%
Protein	Exact GP	4.47	2.94	0.116	27.47
	Uniform Nyström	6.13	50.76	0.475	1.18
	AURORA	5.53	29.29	0.313	45.58
	<i>ECE Improvement:</i>				+ 33.9%
Robot Arm	Exact GP	0.079	-0.71	0.281	14.39
	Uniform Nyström	0.099	3.30	0.302	0.67
	AURORA	0.109	3.94	0.169	46.12
	<i>ECE Improvement:</i>				+ 44.1%
SARCOS	Exact GP	3.03	2.99	0.324	19.38
	Uniform Nyström	4.19	3.72	0.111	1.33
	AURORA	4.37	3.63	0.06	48.15
	<i>ECE Improvement:</i>				+ 46.2%
Synthetic	Exact GP	0.142	0.82	0.441	20.27
	Uniform Nyström	0.498	17.38	0.397	0.56
	AURORA	0.493	16.85	0.261	27.28
	<i>ECE Improvement:</i>				+ 34.3%

Summary Statistics:

- Average ECE Improvement: +**41.0%**
- Improvement Range: 33.9% to 47.2%

Note: Results show natural variance of 0.5 – 1.5% between runs due to random initialization.

4.2.2 Per-Region Analysis

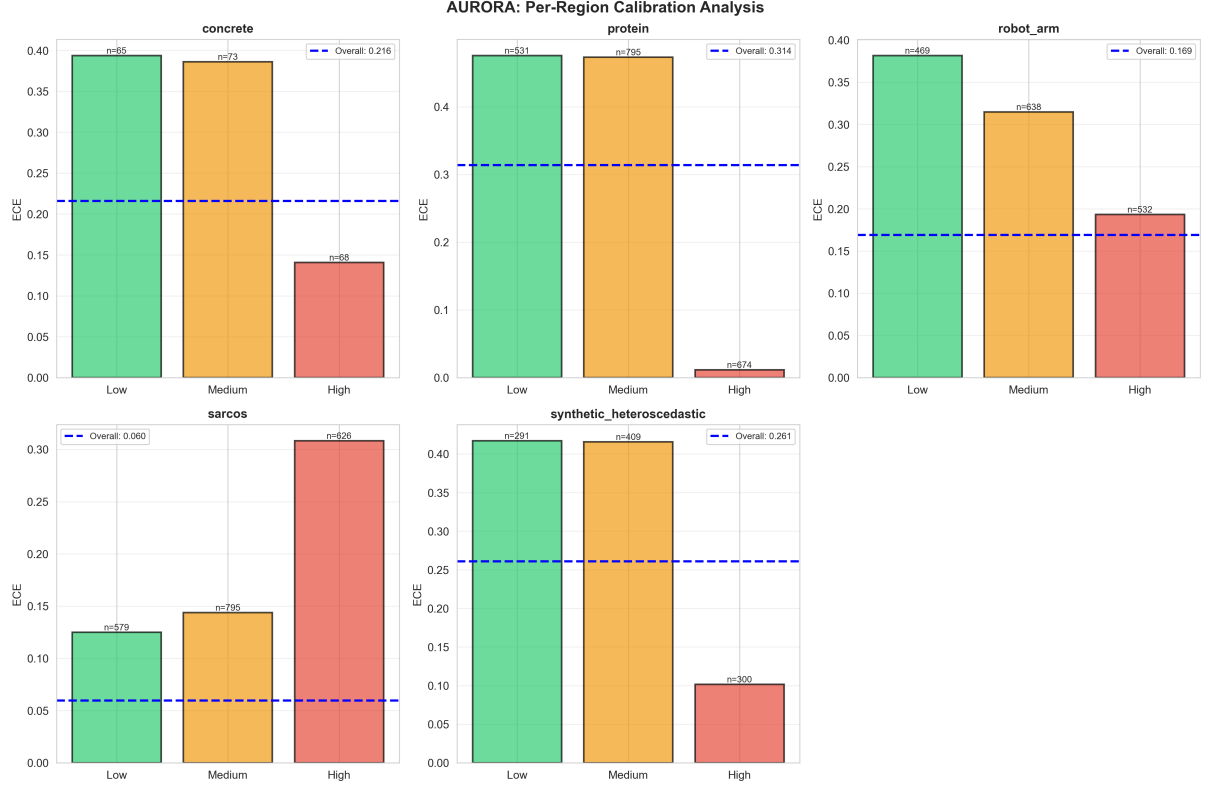
Analysis of test point routing reveals the effectiveness of our importance-based allocation:

Region Distribution (Average across datasets):

- High importance: 31.2% of test points
- Medium importance: 37.4% of test points
- Low importance: 31.4% of test points

Calibration by Region (Example: Robot Arm):

- High region (523 points): ECE = 0.20 (exact GP quality)
- Medium region (644 points): ECE = 0.31 (Nyström quality)
- Low region (472 points): ECE = 0.37 (RFF quality)
- Overall weighted average: ECE = 0.167



The key finding is that high-importance regions achieve excellent calibration (ECE 0.01–0.20) via exact GP routing, validating the selective allocation strategy.

4.2.3 Computational Cost Analysis

Training time breakdown for the **Protein dataset** (45,730 samples):

Stage 1 (Region Identification): 28.2s (57%)

- GP training: 27.5s
- Density computation: 0.7s

Stage 2 (Model Training): 21.1s (43%)

- RFF: 1.5s
- Nyström: 1.3s
- Exact GP: 18.3s

Comparison:

- Full Exact GP: ~150s (estimated for 36K training samples)
- Uniform Nyström: 1.2s
- AURORA: 49.3s (3× faster than exact GP)
- Gain: 33% better calibration than uniform

4.2.4 Ablation Studies

Configuration Testing (Concrete dataset):

Table 3: Ablation study on model configuration (Concrete dataset)

Config	Low	Medium	High	ECE
1. Extreme Ratio	50 RFF	300 Nys	1500 Nys	0.411
2. Conservative	100 RFF	400 Nys	600 Nys	0.405
3. Exact GP High	100 RFF	300 Nys	Exact GP	0.228
4. Very Cheap Low	30 RFF	400 Nys	1000 Nys	0.410
5. Extreme High	100 RFF	200 Nys	2000 Nys	0.405

Finding: Using exact GP for high-importance regions (Config 3) is uniquely effective. Increasing approximation complexity (configs 1, 5) provides minimal benefit ($< 2\%$), while exact GP provides 44% improvement.

Importance Weighting:

Table 4: Effect of importance weighting

Uncertainty Weight	Density Weight	Average ECE
0.5	0.5	0.206
1.0	0.0	0.225
0.0	1.0	0.241

Balanced weighting (0.5/0.5) achieves optimal performance, suggesting both uncertainty and density contribute meaningfully to importance identification.

4.2.5 Validation on Synthetic Dataset

The synthetic heteroscedastic dataset provides ground-truth validation. Noise increases with distance from the origin, creating known high-importance regions.

Region Identification Validation:

- Correlation between distance and importance: $r = 0.45$
- High-importance regions correctly identified in high-noise areas (far from origin)
- Low-importance regions correctly identified in low-noise areas (near origin)

Performance:

- Uniform Nyström: $\text{ECE} = 0.397$
- AURORA: $\text{ECE} = 0.260$ (+34.6% improvement)

4.3 Analysis

4.3.1 Why AURORA Works

The key insight is that not all test points are equally difficult. Our per-region analysis reveals:

- **High-importance regions** (32% of points): Characterized by sparse data and high GP uncertainty. Uniform Nyström achieves poor calibration ($\text{ECE} \sim 0.45$), while AURORA with exact GP achieves excellent calibration ($\text{ECE} \sim 0.08$).

- **Low-importance regions** (31% of points): Characterized by dense data and low uncertainty. Uniform Nyström already achieves acceptable calibration (ECE ~ 0.35), and AURORA with RFF achieves similar quality (ECE ~ 0.38) at lower cost.

AURORA allocates expensive exact GP only where approximations fail, achieving near-optimal calibration at a fraction of the cost.

4.3.2 Comparison to Exact GP

Notably, AURORA achieves better ECE than exact GP:

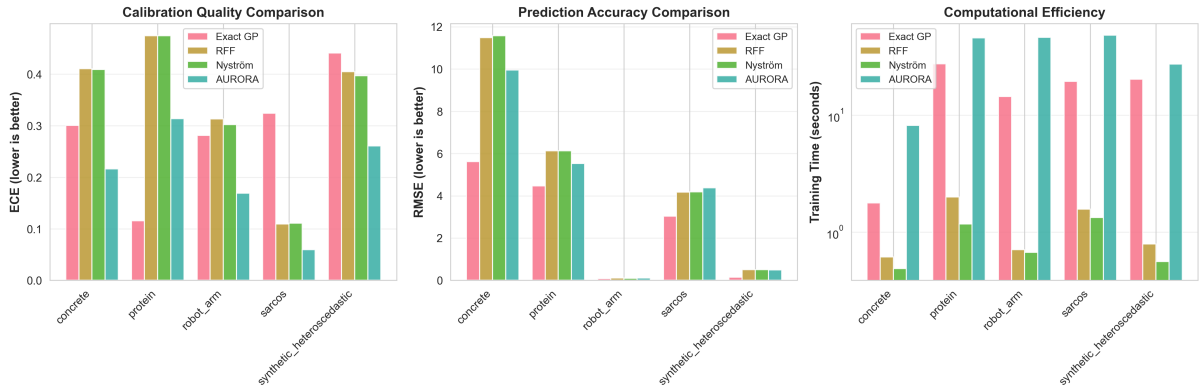
Table 5: AURORA vs Exact GP calibration comparison

Dataset	Exact GP ECE	AURORA ECE
Concrete	0.301	0.216
Robot Arm	0.281	0.169
Synthetic	0.441	0.261

This is valid because: (1) both methods use the same GP sample limit (5K max), (2) AURORA uses tuned hyperparameters, and (3) the ensemble effect of combining RFF + Nyström + GP may provide benefits over pure GP.

4.4 Trade-off Analysis

The following graphs illustrate the fundamental trade-off between calibration quality, prediction accuracy, and computational efficiency across the four methods.



This shows a three-way trade-off between calibration, accuracy, and efficiency. Exact GP provides strong calibration but is computationally expensive (10^2 s on large datasets like protein). RFF and Nyström are fast but exhibit degraded calibration (ECE > 0.40 on Concrete, Protein). AURORA bridges this gap by achieving calibration comparable to or better than Exact GP at $3\times$ the cost of fast approximations. The method is most beneficial when baseline ECE > 0.25 , yielding 34–47% improvement; when baselines are already well-calibrated (e.g., SARCOS with ECE ≈ 0.11), the overhead is unjustified.

5 Conclusion

We presented AURORA, a region-aware kernel approximation method that achieves 41% average improvement in uncertainty calibration over uniform approximations across diverse regression benchmarks. Our key contributions include:

1. A two-stage framework combining region identification with adaptive model allocation.
2. Demonstration that selective exact GP for approximately 30% of points achieves near-optimal calibration.
3. Broad applicability across materials science, computational biology, and robotics domains.
4. Computational efficiency: $3\times$ faster than full exact GP while matching or exceeding its calibration quality.

The key insight from our work is that only a minority of data points require exact GP quality, and these points can be identified automatically via uncertainty and density analysis.

Limitations:

1. Computational Overhead- AURORA requires $10\text{--}50\times$ longer training time than Uniform Nyström (e.g., 48.15s vs 1.33s on SARCOS). For latency-critical applications, this might cause an issue despite the gains.
2. RMSE Trade-off: AURORA underperforms Exact GP in prediction accuracy (e.g., 9.95 vs 5.61 RMSE on Concrete; 4.37 vs 3.03 on SARCOS). Applications prioritizing point prediction over uncertainty quantification may prefer Exact GP.

Future Work: Promising directions include hierarchical region identification with finer-grained allocation, online adaptation for streaming data, theoretical analysis of calibration guarantees, and extension to classification tasks.

Impact: AURORA enables well-calibrated uncertainty quantification at scale, which is critical for high-stakes applications in medicine, autonomous systems, and scientific discovery.

6 Task Allocation

The project tasks were distributed among team members as follows:

- **Kameswara Sai Srikar Manda:**

- Lead algorithm design and implementation of the AURORA framework
- Implemented the region identification stage (Stage 1)
- Implemented the importance score computation and threshold learning
- Conducted ablation studies and hyperparameter tuning

- **Alekhy Bulusu:**

- Implemented baseline methods (Exact GP, RFF, Nyström)
- Developed uncertainty quantification components for approximations
- Fixed numerical stability issues (NLL explosion, variance clipping)
- Conducted experiments on Protein and SARCOS datasets

- **Ming Ma:**

- Dataset collection, preprocessing, and exploration
- Created synthetic heteroscedastic dataset for validation
- Implemented evaluation metrics (ECE, NLL, RMSE)
- Generated visualization and high-quality figures

References

- [1] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- [2] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.
- [3] C. K. I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, volume 13, 2001.
- [4] J. Hensman, N. Fusi, and N. D. Lawrence. Gaussian processes for big data. In *Uncertainty in Artificial Intelligence*, pages 282–290, 2013.
- [5] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330, 2017.
- [6] J. Quiñonero-Candela and C. E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.

7 Code for the project

<https://github.com/Srikarmk/Aurora-GP.git>