# Combining Thermodynamics-based Model of the Centrifugal Compressors and Active Machine Learning for Enhanced Industrial Design Optimization

Shadi Ghiasi [1] Guido Pazzi [2] Concettina Del Grosso [2] Giovanni De Magistris [1] Giacomo Veneri [1]

## Abstract

The design process of centrifugal compressors requires applying an optimization process which is computationally expensive due to complex analytical equations underlying the compressor's dynamical equations. Although the regression surrogate models could drastically reduce the computational cost of such a process, the major challenge is the scarcity of data for training the surrogate model. Aiming to strategically exploit the labeled samples, we propose the *ActiveCompDesign* framework in which we combine a thermodynamics-based compressor model (i.e., our internal software for compressor design) and Gaussian Process- based surrogate model within a deployable Active Learning (AL) setting. We first conduct experiments in an offline setting and further, extend it to an online AL framework where a real-time interaction with the thermodynamics-based compressor's model allows the deployment in production. *ActiveCompDesign* shows a significant performance improvement in surrogate modeling by leveraging on uncertainty-based query function of samples within the AL framework with respect to the random selection of data points. Moreover, our framework in production has reduced the total computational time of compressor's design optimization to around 46% faster than relying on the internal thermodynamics-based simulator, achieving the same performance.

## 1. Introduction and Related Works

Centrifugal compressors' design is an extensive computational process as it requires the optimization of numerous design variables which are the starting point of software simulations of complex dynamical equations (Ju et al., 2021).

While engineering-powered software simulations are reliable solution to the end user due to the established technology, they have more complex formulations due to higher inter-dependency of system variables (Garg et al., 2010).

Surrogate modeling with Machine Learning (ML) models for computer simulations enables reducing the computational cost and time required for a design simulation while maintaining a desired performance in industrial applications (Bicchi et al., 2022; Owoyele et al., 2022; Kim et al., 2010). However, generating sufficient data points for training ML models is a daunting task since it requires running extensive software simulations. Therefore, without any strategic sampling, the possibility to explore a larger design space is limited.

Under such circumstances, the Active Learning (AL) strategy is a powerful framework to alleviate the problem of high quality annotation scarcity (Settles, 2012). AL is a ML technique that allows the model to interact with an oracle by queering the most important data for learning (Monarch, 2021). In industrial applications, AL can make the most of resources by significantly reducing the amount of labeled data for training ML models (Brevault et al., 2022).

Utilizing ML surrogate models in the industrial simulation design setting has been explored by previous research. Kim et al. implement surrogate modeling for optimization of a centrifugal compressor impeller (Kim et al., 2010). However, without any strategic sampling, the research is done for a limited design space. *AutoML-GA* (Owoyele et al., 2022) is an application of an automated machine learning-genetic algorithm coupled with computational fluid dynamics simulations for rapid engine design optimization. Chabanet et al. (Chabanet et al., 2021) apply AL in Industry 4.0 context. Moreover, Murugesan et al. (Murugesan et al., 2022) propose an AL framework for estimating the operating point of a Modular Multi Pump used in energy field. Wang et al. (Wang & Nalisnick, 2022) apply AL for multilingual finger spelling corpora. Finally, see also (Reker, 2019) for some practical considerations for active ML in drug discovery. However, an AL based framework has not been explored

[1] Artificial Intelligence Team - Baker Hughes, Florence, Italy
[2] Software Engineering Team - Baker Hughes, Florence, Italy. Correspondence to: Shadi Ghiasi <shadi.ghiasi@bakerhughes.com>.

*Accepted after peer-review at the 1st workshop on Synergy of Scientific and Machine Learning Modeling*, SynS & ML ICML, Honolulu, Hawaii, USA. July, 2023. Copyright 2023 by the author(s).

for design optimization of centrifugal compressors. Moreover, most research have focused on offline evaluation of AL strategies, while, in industrial settings the data is acquired in real-time (Cacciarelli et al., 2022) and a deployable streaming based AL framework is needed.

In this study, we present the *ActiveCompDesign* framework for deployable AL based design optimization of centrifugal compressors. We leverage on Gaussian Processes (GPs) as deep surrogates of centrifugal compressor dynamics coupled with the AL strategy with a design goal to reach the optimal power absorbed by the machine. We perform extensive computer simulations using our internal thermodynamics-based model integrated with an optimization algorithm to generate sufficient data samples for surrogate model training. We then use this data to perform an offline AL algorithm with GP surrogates as a proof of concept. We further deploy our framework through an online AL simulation environment in which the thermodynamics-based model and the ML-based model interact in real-time using a stream-based AL strategy. Our framework is currently in production. To the best of our knowledge, no other study have combined compressor's thermodynamics-based models and ML to propose a production-ready AI enhanced design optimization framework for design optimization of centrifugal compressors.

## 2. *ActiveCompDesign* framework

### 2.1. Problem definition

We consider the optimization of the boil off gas process of centrifugal compressors where the aim is to minimize the absorbed power by the machine. This is currently done based on a design optimization process on an internal thermodynamics-based simulator. Since the analytical equations underlying this physical process is complex and computationally expensive, we aim to strategically run the thermodynamics-based simulator during the optimization process with the minimum number of queries to reach the desired power.

We do this within the *ActiveCompDesign* framework by integrating a regression surrogate model of the physical process throughout the optimization process to benefit from a faster calculation of the system's response. However, due to the uncertainties produced by the surrogate model, we want to rely on the actual physical process equations to obtain reliable outputs when the ML surrogate model produces high uncertainties in prediction. In the following subsections we provide formulation for surrogate modeling of the design optimization process and the implemented offline and online AL framework.

### 2.2. Surrogate modeling and active learning

Let $\mathbf{x} \subset \mathbb{R}^{d_{\text{in}}}$ be the input set where $d$ is the number of design parameters chosen by the expert engineer and $\mathbf{y} \subset \mathbb{R}^{d_{\text{out}}}$ is the desired output. A collection of $N$ number of runs of the thermodynamics-based simulator, will result in a finite number of pairs $\mathcal{D}_{train} = (\mathbf{x}_n, \mathbf{y}_n)_{n=1}^{N}$ which is considered as the training data. These pairs represent sampled inputs and outputs of a complex analytical function $\mathbf{y} = f(\mathbf{x})$ which is encoded in the simulator. The aim of surrogate modeling is to estimate a function $\widehat{f} : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$ which should be as close as possible to the true function $f$. We aim to condition the parameters of $\widehat{f}(\theta)$ through $\widehat{f}(x, \theta|\mathcal{D}_{train}) = \widehat{y}, \forall x \in \mathcal{B} : \widehat{y} \sim y$ where $\mathcal{B}$ is the bounded subspace in $\mathbb{R}$. In design optimization we look for those values of $x$ leading to the minimum $y$. Therefore, by integrating the surrogate model into the optimization algorithm we aim to:

$$min| \sim y \text{ subject to} \widehat{f}(x, \theta|\mathcal{D}_{train}) = \widehat{y}, \forall x \in \mathcal{B} : \widehat{y} \sim y \tag{1}$$

In the *ActiveCompDesign* framework, our design parameters ($x$) are a set of 12 flow coefficient rates corresponding to different compression stages. Each coefficient is bounded between a minimum and maximum value designed by the expert engineer. The output is the total absorbed power ($\mathcal{P}$) calculated by the thermodynamics-based simulator. Therefore, a single objective optimization algorithm is performed to reach a desired $\mathcal{P}$.

### 2.3. Offline *ActiveCompDesign* framework

We gather data by running N number of runs of the thermodynamics-based model during an optimization process selected by the expert engineers to obtain a baseline minimum power. Given this collected dataset ($\mathcal{D}_t^{Start}$) comprising of a set of flow coefficents as $x$ and power as $\mathcal{P}$, we train a regressor $\widehat{f}^{Start}$ on a small pool of labeled data ($\mathcal{D}_t^{Pool}$). Throughout the AL process the regressor and the original dataset get updated resulting in $\widehat{f}_{Up}$ and $\mathcal{D}_t^{Up}$. The selection of the new observations to be labeled by the simulator to be considered for the new training dataset is obtained using query strategies formulated based on the regressor.

We particularly choose GPs (Williams & Rasmussen, 2006) as the surrogate regression model thanks to their high performance in mapping input-output relationship in our study and their Bayesian structure (A list of regression models tested in our data set and a comparison of their performance is in the Appendix). A GP is defined by its mean function $\mu(x)$ and a Kernel function computing the covariance function between datapoints $K(x_i, x_j)$, therefore $\widehat{f} \sim \mathcal{GP}(\mu(x_i), K(x_i, x_j))$.

Within the GP regression framework we are able to compute

**Algorithm 1** Online *ActiveCompDesign* framework

```
Input: Flow coefficients ($X$), Bounded space of flow coefficients ($\mathcal{B}$), Number of iterations for pre-training ($N_{PT}$), Number of total iterations ($N_{tot}$)
Output: Power absorbed by the machine ($\mathcal{P}$)
Initialize $X$= arbitrary initial flow coefficient values.
for $i = 1$ to $N_{PT}$ do
    Get the input ($X$) from the optimizer within ($\mathcal{B}$).
    Compute the power $\mathcal{P}$ based on thermodynamics-based model ($\mathcal{P}_{Phys}$).
    Create $\mathcal{D}^{Pretrain}$
    Train the pre-trained $\mathcal{GP}$ model ($\mathcal{GP}_{pt}$) on this training set.
    Compute the uncertainty threshold $Ub$ based on $\mathcal{GP}$ posterior variances of previous samples.
end for
for $i = N_{PT} + 1$ to $N_{tot}$ do
    Get the input ($X$) from the optimizer within ($\mathcal{B}$).
    Compute the power calculated from GP ($\mathcal{P}_{GP}$), uncertainty ($\mathcal{P}_{GP}$) from $\mathcal{GP}_{pt}$.
    if Uncertainty ($\mathcal{P}_{GP}$) $< Ub$ then
        Return $\mathcal{P}$=$\mathcal{P}_{GP}$
        Update $Ub$.
    else
        Return $\mathcal{P}$=$\mathcal{P}_{Phys}$.
        Update $\mathcal{GP}_{pt}$ with the new data point.
        Update $Ub$.
    end if
end for
```

*Table 1.* Computational time of the proposed framework compared to the baseline thermodynamics-based-simulator. N shows the number of iterations for pre-training and total number of runs are the number of iterations for *ActiveCompDesign* to achieve the minimum baseline power.

| | Number of runs | Pre-training time (s) | Total time (s) |
|---|---|---|---|
| Thermodynamics model (baseline) | 4000 | - | 109080 |
| *ActiveCompDesign* ($N_{PT}$=50) | 200 | 2340 | 7180 |
| *ActiveCompDesign* ($N_{PT}$=100) | 160 | 4680 | 5040 |
| *ActiveCompDesign* ($N_{PT}$=150) | 120 | 7020 | 6280 |

the posterior mean and posterior variance for the prediction of each sample. To perform AL with GP we select the next training sample based on the maximum variance (Kapoor et al., 2007; Zhao et al., 2021; Yue et al., 2020; Zimmer et al., 2018). For implementation of offline *ActiveCompDesign* we rely on the *ModAL* library (Danka & Horvath, 2018).

### 2.4. Online *ActiveCompDesign* framework

With this approach we design a simulation environment where the thermodynamics-based simulator, the optimizer and the surrogate model can interact in real time after each data streaming.

The primary difference between the online and offline framework is that the labels of new observations can non longer be queried in a pool based way. With real time streaming of data points an instant decision should be made to whether label the data points or to discard the sample for learning.

Since our optimization goal is to reach minimum power ($\mathcal{P}$) with less expensive computational effort but keeping the reliability of the output, we rely on the GP surrogate model only as an alternative model in case the uncertainty of the prediction is high. To this end, we consider a maximum number of iterations ($N_{PT}$) for the optimizer and the thermodynamics-based simulator to interact to generate an initial labeled dataset $\mathcal{D}^{Pretrain} = (\mathbf{x}_n, \mathbf{y}_n)_{n=1}^{N_{tot}}$. We then train a surrogate model based on GPs and consider the mean of the posterior variances of the predictions as the uncertainty threshold ($Ub$) for the next iteration. We examine if this replicated simulation is able to achieve comparable performance in suggesting optimal design parameters for obtaining the minimum power. Our thermodynamics-based simulator and the code for the framework in this paper are proprietary. More details of the online framework is reported in algorithm 1.

## 3. Results and Discussion

### 3.1. Offline *ActiveCompDesign* framework

To generate the dataset for training and evaluating the GP surrogate model, we perform a Bayesian optimization using GPs (from the *Scikit-optimize library (Louppe, 2017)*) on top of our thermodynamics-based simulator of compressor design. This optimization process has led to 4000 runs in 12 dimensional input parameter space which constructs our labeled dataset.

We select the surrogate GPs's hyper-parameters by performing a grid search hyper-parameter selection based on cross validation performance (Bergstra et al., 2011). The best performing model is achieved with the *Matern* kernel with length scale of 0.75 and $\nu$ of 0.5 (refer to (Williams & Rasmussen, 2006) for formulation of the *Matern* kernel).

We perform model training with AL and compare the results with uniformly random acquisition of the samples as the baseline. Figure 1 reports the regression performance metrics (i.e., Root Mean Squared Error (*RMSE*), *R-squared*, Mean Absolute Percentage Error (*MAPE*) and *Max error*

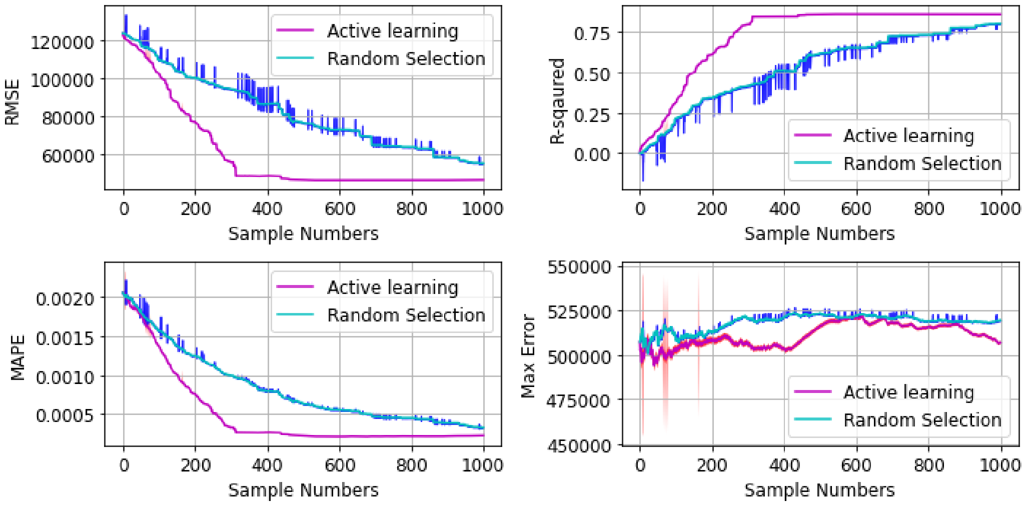

*Figure 1.* Comparison of performance metrics of surrogate modeling with Gaussian Process regression model by selecting the samples with random acquisition with respect to active learning framework. The vertical bars for each sample show the standard deviation of the performance metric as a result of uncertainty for prediction.

as more samples are queried for training based on random selection and AL. We observe that the uncertainty-based query strategy leads to a much greater decrease of *RMSE* and *MAPE* metrics with less samples compared to the random sampling. This has lead to a clear improvement of error achieving almost the full-dataset performance by relying on only around 30% of labeled data points for training. Moreover, the model's goodness of fit quantified by *R-squared* has a faster increase using the AL strategy compared to the random selection.

As through GPs, we are able to obtain the standard deviation of prediction ($\sigma$), we also consider the regression performance for $y$+$\sigma$ and $y$-$\sigma$ where $y$ is the predicted value for each sample. As the plots in Figure 1 show, there are high variations in regression performances with random selection, while with AL, this variation exists only for few initial samples and it disappears as the samples strategically grow in number.

### 3.2. Online *ActiveCompDesign* framework

We evaluate the results of the deployed streaming-based AL framework based on the total number of runs required for achieving the baseline performance and the computational cost associated with it. In this regard, we consider different simulations where in each simulation $N_{PT}$ number of initial iterations is needed to pre-train the surrogate model. We compare these results in Table 1. The baseline model is the thermodynamics-based simulator where it achieved a desired minimum power through 4000 iterations lasting for around 30 hours. These results show a trade-off between the number of runs to interact with thermodynamics-based model and the total number of iterations for the whole framework to achieve the desired $\mathcal{P}$. For the *ActiveCompDesign* simulators, the highest computational cost is the pre-training cost where an interaction with oracle is needed. The simulator with 100 iterations for pre-training with a total number of 160 runs have taken the least total time of simulation with an improvement of around 46% decrease in total computational time compared to the thermodynamics-based model. All our experiments have been performed using *Nvidia- DGX1* with *8x Tesla P100* GPUs and 20-core dual *Intel* CPUs.

## 4. Conclusion

The benefits of combining active ML methods with physical models underlying compressor's dynamics are large for design optimization applications including faster computations and more accurate design solutions. However, the trade-off between performance and computational power has to be carefully evaluated for the specific design application. Moreover (see also (Pardakhti et al., 2021)), the adoption of AL methods poses significant challenges in many practical applications, such as lack of data, discontinuous space of exploration and measurement error.

The results obtained from offline and online *ActiveCompDesign* show that integrating ML in compressor's simulators is viable for production ready application on the energy sector. Indeed, our framework is currently working in a production environment. For future research we will expand this framework for wider design optimization applications and improve the monitoring of the deployed model.

## 5. Impact Statement

The inclusion of ML models into an internal thermodynamics-based model in our study offers a significant positive impact on various aspects of design optimization process in the turbo-machinery industry, including performance improvement, code optimization and enhancement of user experience. However, it is important to address the negative impacts such as privacy concerns and the potential inaccuracies in model prediction generated by dataset distribution shift and other factors.

## References

Bergstra, J., Bardenet, R., Bengio, Y., and Kégl, B. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.

Bicchi, M., Biliotti, D., Marconcini, M., Toni, L., Cangioli, F., and Arnone, A. An AI-based fast design method for new centrifugal compressor families. *Machines*, 10(6): 458, 2022.

Brevault, L., Balesdent, M., and Valderrama-Zapata, J.-L. Active learning strategy for surrogate-based quantile estimation of field function. *Applied Sciences*, 12(19): 10027, 2022.

Cacciarelli, D., Kulahci, M., and Tyssedal, J. Online active learning for soft sensor development using semi-supervised autoencoders. *arXiv preprint arXiv:2212.13067*, 2022.

Chabanet, S., El-Haouzi, H. B., and Thomas, P. Coupling digital simulation and machine learning metamodel through an active learning approach in industry 4.0 context. *Computers in Industry*, 133:103529, 2021.

Danka, T. and Horvath, P. modal: A modular active learning framework for python. *arXiv preprint arXiv:1805.00979*, 2018.

Garg, S. K., Buyya, R., and Siegel, H. J. Time and cost trade-off management for scheduling parallel applications on utility grids. *Future Generation Computer Systems*, 26(8):1344–1355, 2010.

Ju, Y., Liu, Y., Jiang, W., and Zhang, C. Aerodynamic analysis and design optimization of a centrifugal compressor impeller considering realistic manufacturing uncertainties. *Aerospace Science and Technology*, 115:106787, 2021.

Kapoor, A., Grauman, K., Urtasun, R., and Darrell, T. Active learning with gaussian processes for object categorization. In *2007 IEEE 11th international conference on computer vision*, pp. 1–8. IEEE, 2007.

Kim, J.-H., Choi, J.-H., and Kim, K.-Y. Surrogate modeling for optimization of a centrifugal compressor impeller. *International Journal of Fluid Machinery and Systems*, 3 (1):29–38, 2010.

Louppe, G. Bayesian optimisation with scikit-optimize. In *PyData Amsterdam*, 2017.

Monarch, R. M. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

Murugesan, M., Goyal, K., Barriere, L., Pasquotti, M., Veneri, G., and De Magistris, G. Deep surrogate of modular multi pump using active learning. *arXiv preprint arXiv:2208.02840*, 2022.

Owoyele, O., Pal, P., Vidal Torreira, A., Probst, D., Shaxted, M., Wilde, M., and Senecal, P. K. Application of an automated machine learning-genetic algorithm (AutoML-GA) coupled with computational fluid dynamics simulations for rapid engine design optimization. *International Journal of Engine Research*, 23(9):1586–1601, 2022.

Pardakhti, M., Mandal, N., Ma, A. W., and Yang, Q. Practical active learning with model selection for small data. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1647–1653. IEEE, 2021.

Reker, D. Practical considerations for active machine learning in drug discovery. *Drug Discovery Today: Technologies*, 32:73–79, 2019.

Settles, B. Active learning. *Synthesis lectures on artificial intelligence and machine learning*, 6(1):1–114, 2012.

Wang, S. and Nalisnick, E. Active learning for multilingual fingerspelling corpora. *Adaptive Experimental Design and Active Learning in the Real World*, 2022.

Williams, C. K. and Rasmussen, C. E. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Yue, X., Wen, Y., Hunt, J. H., and Shi, J. Active learning for gaussian process considering uncertainties with application to shape control of composite fuselage. *IEEE Transactions on Automation Science and Engineering*, 18 (1):36–46, 2020.

Zhao, G., Dougherty, E., Yoon, B.-J., Alexander, F., and Qian, X. Efficient active learning for Gaussian process classification by error reduction. *Advances in Neural Information Processing Systems*, 34:9734–9746, 2021.

Zimmer, C., Meister, M., and Nguyen-Tuong, D. Safe active learning for time-series modeling with gaussian processes. *Advances in neural information processing systems*, 31, 2018.

# A. Appendix

## A.1. General overview of the *ActiveCompDesign* framework

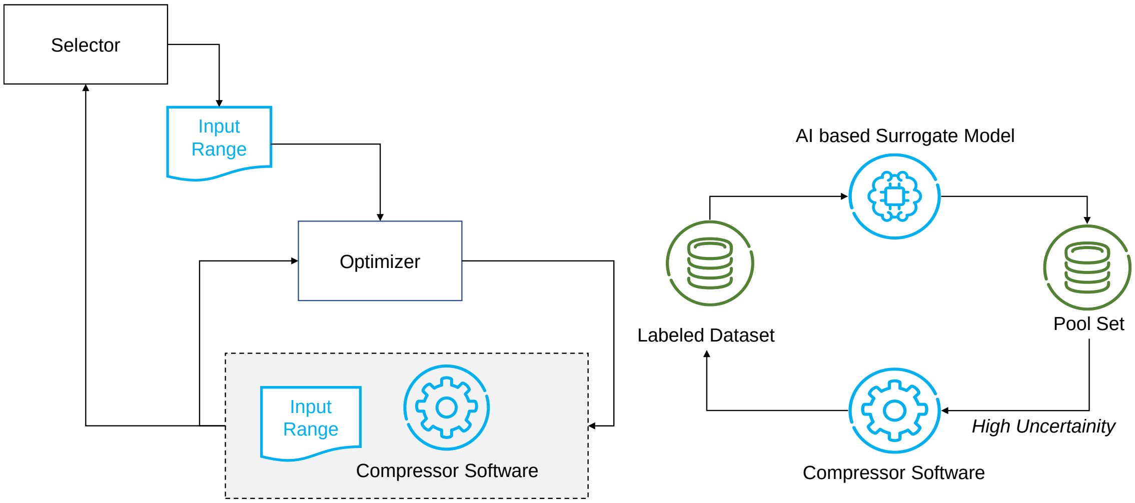

*Figure 2.* General pipeline of *ActiveCompDesign* framework.

## A.2. Surrogate modeling selection

In order to obtain an AI based surrogate model of the compressor's designer software, we rely on state-of-the-art supervised regression models to map the input-output relationship. Moreover, we select those regression models where we are able to obtain a quantification of the uncertainty of the prediction. Here a comparison of the performance of all tested models is listed 2. For random forests, gradient boosting and extra tree regressors we consider 100 numbers of estimators for training. For GPs we use a Matern kernel with parameters described in the text. We observe that GPs have achieved the best performance among all the models.

*Table 2.* Model development performance.

| Model | *RMSE* | *Max Error* | *R Squared* | *MAPE* |
|---|---|---|---|---|
| Random Forest | 18937.12 | 112618.01 | 0.84 | 0.0003 |
| Gaussian Process | 4412.71 | 53286.78 | 0.99 | 0.00002 |
| Gradient Boosting | 16919.41 | 50417.80 | 0.87 | 0.0002 |
| Extra Tree Regressor | 18413.44 | 106056.33 | 0.85 | 0.0003 |