ChemComm

# SelTox: Discovering the Capacity of Selectively Antimicrobial Nanoparticles for Targeted Eradication of Pathogenic Bacteria



| | |
|---|---|
| Journal: | *ChemComm* |
| Manuscript ID | Draft |
| Article Type: | Communication |
| | |
| Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online. | |
| Jyakhwo_RSC_ChemComm.zip | |

Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

# SelTox: Discovering the Capacity of Selectively Antimicrobial Nanoparticles for Targeted Eradication of Pathogenic Bacteria†

Susan Jyakhwo,[a] Valentina Bocharova,[a] Nikita Serov,[a] Andrei Dmitrenko,[a] Vladimir V. Vinogradov,[a*]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

This study proposes an innovative approach to discover selectively antimicrobial nanoparticles (SANPs) for targeted eradication of pathogenic bacteria while minimizing harm to non-pathogenic ones, using machine learning reinforced genetic algorithm. As a proof of concept, CuO SANPs were identified for the targeted eradication of Klebsiella pneumoniae.

Microbial infections, caused by bacteria, viruses, fungi, or parasites, can lead to diverse human diseases affecting various parts of the body.[1,2] Antibiotics, designed to target and eradicate infection causing pathogens, have played a pivotal role in modern medicine saving numerous lives. However, extensive use of antibiotics escalates threat of multi-drug resistant pathogen, highlighting the urgent need of alternative antimicrobial strategies.[3,4]

Inorganic nanoparticles (NPs) have emerged as promising alternatives to existing antibiotic drugs due to their ability to eradicate microbial infection and inhibit microbial growth. For example, metal and metal oxide NPs, such as silver (Ag)[5], gold (Au)[6], copper oxide (CuO)[7], and zinc oxide (ZnO)[8] NPs implement their antibacterial activity by disrupting bacterial metabolism[9], damaging cell membranes, releasing reactive oxygen species (ROS), and preventing bacterial adhesion and biofilm formation.[10,11] NPs interact with bacterial cells (via receptor-ligand,[12] hydrophobic interactions[13] and more) and affect key cellular components, inducing oxidative stress and alteration in gene expression.[14,15] However, it is noteworthy that due to diverse mechanisms of actions, identification of selectively antimicrobial nanoparticles (SANPs) against specific bacterial strains remains challenging.

The process of identification of SANPs encompasses intricate, arduous and time-consuming synthesis, characterization testing and validation steps.[16,17] To address these experimental challenges, machine learning (ML) methods have emerged as fast and cost-effective tools for predicting the antimicrobial activity of inorganic NPs on various microorganisms.[18] These ML techniques leverage big datasets and advanced algorithms to identify complex patterns, enabling the prediction of NP properties.[19,20] While researchers have implemented various ML algorithms for predicting antimicrobial activity,[18,21] to the best of our knowledge, no study addressed the discovery of SANPs against specific microorganisms. The SANPs are characterized by their ability to exclusively possess toxicity for specific strain of microbes while at the same time they do not harm other strains. We implemented ML reinforced genetic algorithm (GA) for screening SANPs that can eradicate pathogenic bacteria with minimal harm to beneficial ones and discovered the rod shaped CuO NPs with minimal bactericidal concentration (MBC) value of 62.52 $\mu g$/ml for pathogenic bacteria *Staphylococcus aureus* and 455.37 $\mu g$/ml for

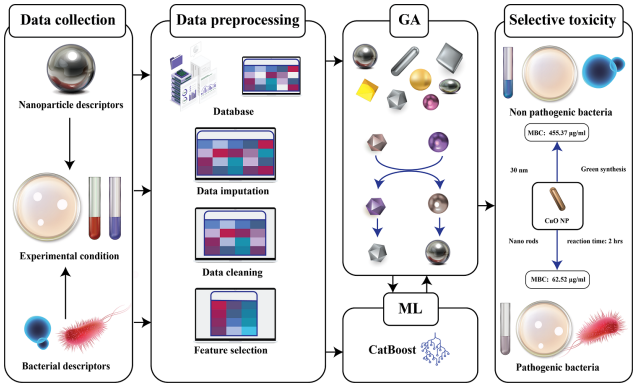



Fig. 1 Schematic representation of discovery of selectively antimicrobial nanoparticles (SANPs) using machine learning (ML) reinforced genetic algorithm (GA).

[a] *International Institute "Solution Chemistry of Advanced Materials and Technologies", ITMO University, Saint-Petersburg 191002, Russian Federation*
[*] *Corresponding author; Email: vinogradov@scamt-itmo.ru*
† Electronic Supplementary Information (ESI) available: A separate file is provided with all the supporting information. See DOI: 00.0000/00000000.

non-pathogenic *Bacillus subtilis* (see detailed workflow of ML reinforced GA in Figure 1).

To achieve this, two comprehensive but distinct datasets were compiled encompassing physicochemical properties of NPs, experimental parameters, microbial taxonomic and biochemical properties (see the Supplementary Information (SI) section for the details on the data collection process). The first dataset (hereafter referred to as the MC dataset) includes 789 samples with 80 features and was used to predict minimal concentration of NPs required to eradicate or inhibit microbial growth. Similarly, the second dataset (hereafter, the ZOI dataset) with 920 samples and 80 features was compiled to predict the Zone of inhibition (ZOI).

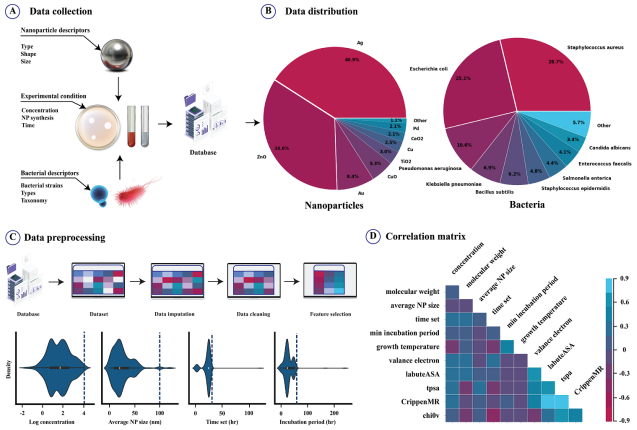



Fig. 2 Data description. A) Data collection process with nanoparticles, experimental and bacterial descriptors. B) Distribution of nanoparticle and bacterial type in MC dataset. C) Data preprocessing and distribution of log concentration, average NP size, time set and incubation period with outlier cutoff threshold as blue dot line in MC dataset. D) Correlation matrix representing Pearson correlation coefficient between pairs of numerical parameters present in the MC dataset.

Data collection step was followed by standard preprocessing steps: removal of the samples with missing values, outliers, and duplicates, to ensure the adequate performance of ML models (Figure 2c). Moreover, the low variance features were discarded, and the highly correlated features were filtered out using the criteria $|r| > 0.95$ for Pearson correlation coefficient (Figure 2d). After the preprocessing steps, the dataset was reduced to 489 samples of 25 features for the MC dataset and 609 samples of 27 features for the ZOI dataset (the details on the data distribution and preprocessing of MC and ZOI datasets are provided in SI, Figure S1-S4, Table S1-S4).

Subsequently, we identified CatBoost Regressor, and XGB Regressor as best performing regression models by evaluating the performance of 43 different supervised regression models on raw and preprocessed datasets. (Figure 3a; see details regarding model performance in SI, Figure S5 & S6, Table S5-S8). We optimized both models and observed similar improvement in their performance with a slightly better metrics for CatBoost regressor with the optimized 10-fold CV scores of 0.82 and 0.84, and with the RMSE scores of 0.46 and 2.41 for the MC dataset and the ZOI dataset, respectively (Figure 3b).

The analysis of critical features aimed at identifying key factors influencing the prediction of NP antimicrobial activity revealed

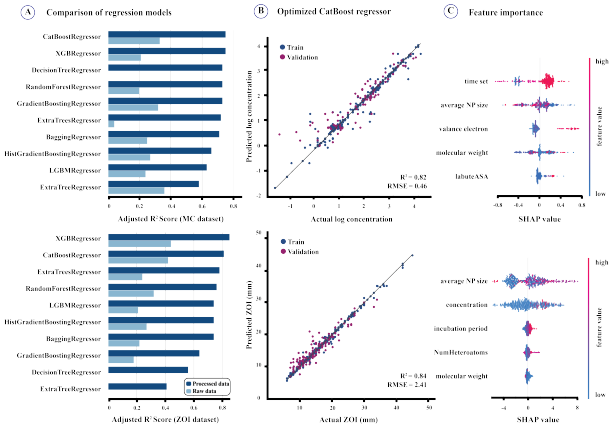



Fig. 3 Machine learning model evaluation and optimization for prediction of antimicrobial activity of nanoparticles. A) Comparison of adjusted $R^2$ score of top 10 regressor models on raw and preprocessed datasets (MC and ZOI). B) Performance of optimized CatBoost regressor on MC dataset ($R^2$ score = 0.82, root mean square error (RMSE) = 0.46) and on ZOI dataset ($R^2$ score = 0.84, RMSE = 2.41). C) Top 5 most important features obtained by SHAP analysis on the MC dataset and the ZOI dataset, respectively.

that model's predictive power is primarily influenced by the fundamental experimental parameters, especially by time, concentration and reaction time, as anticipated.[22] Although NP size is a key parameter, the model couldn't isolate its direct impact as it is also influenced by other size-related factors like surface area, shape, and crystal structure, which are often overlooked in research literature.[7,23,24] Furthermore, other important features, valence electrons and heteroatoms, are linked to ROS generation, a key mechanism in antimicrobial activity.[25,26] This showed that our model is effective and able to interpret NP interactions with microbial systems.

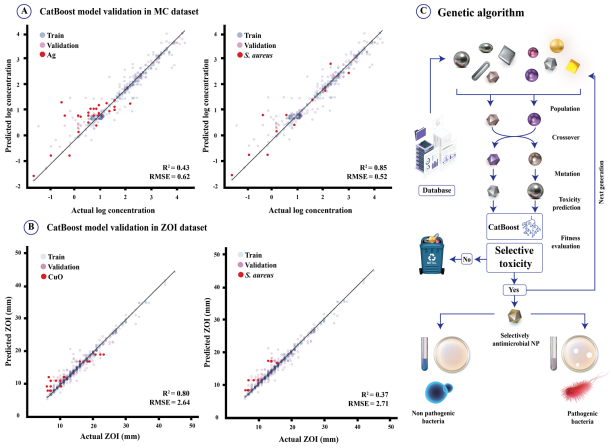



Fig. 4 Evaluation of model performance on various nanoparticles and bacterial strains A) Model performance on Ag NPs and *Staphylococcus aureus* bacteria in MC dataset, and B) Model performance on CuO NPs and *Staphylococcus aureus* bacteria in ZOI dataset. C) Schematic representation of machine learning (ML) reinforced genetic algorithm (GA).

To delve deeper, we evaluated the model's generalization ability on both MC and ZOI datasets and observed $R^2$ score of 0.67

with an RMSE of 0.58 for the MC dataset, and an $R^2$ score of 0.66 with an RMSE of 3.05 for the ZOI dataset, indicating a slight decline in performance. Further analysis involved assessing the model's limitations across various NP types and bacterial strains (Figure 4a & 4b, see details in SI, Figure S7-S10, Table S9 & S10). The most likely reason for the drops in performance is the uneven initial data distribution, where some NP and bacterial types were underrepresented.

Building on our previous research,[27] we developed the ML reinforced GA, a high throughput *in silico* screening approach, combining QSAR and read-across techniques to establish a quantitative connection between the physicochemical properties of NPs and their activity in biological systems.[28,29] As a proof of concept, we applied this platform to identify SANPs for treating pneumonia, which affects approximately 489 million people annually and remains as a major health issue[30] and stands as a leading cause of death in developing countries.[31] Our targeted pathogens were *Staphylococcus aureus* and *Klebsiella pneumoniae* which cause pneumonia. Using the ML reinforced GA on the MC dataset, we identified the CuO NPs (with the main parameters: NPs obtained with green synthesis, nanorods in shape, average NP size of 30 nm, and reaction time of 2 hours) as one of the top SANPs. It showed MBC of 62.52 $\mu$g/ml against *Staphylococcus aureus* while it achieved MBC of 455.37 $\mu$g/ml for *Bacillus subtilis*. So, the CuO SANPs is more toxic against pathogenic *Staphylococcus aureus* compared to non-pathogenic *Bacillus subtilis* with concentration difference of 392.85 $\mu$g/ml. Similarly, we identified ZnO SANPs exhibiting minimal concentration difference of 285.5 $\mu$g/ml against *Klebsiella pneumoniae* (with MBC of 52.5 $\mu$g/ml) compared to *Bacillus subtilis* (with MBC of 338 $\mu$g/ml). In the same way, using the ML reinforced GA on the ZOI dataset, we discovered the $TiO_2$ and Ag SANPs with selective toxicity against *Bacillus subtilis* and *Klebsiella pneumoniae* respectively (see the list of top SANPs in SI, Table S11-S14).

These findings showed that our platform can indeed discover the unique SANPs that feature selective toxicity against some specific pathogenic strains of bacteria, with reduced toxicity on non-pathogenic strains of bacteria. To affirm these findings, we used the model to reproduce the experimental results described in research literature where the toxicity of some specific NPs had been tested for various bacterial strains. Namely, we referred to the study conducted by Pathak et al.[32], which evaluated MIC and MBEC of ZnO NPs of average size 50 nm, and that conducted by Smekalova et al.[33], which evaluated MIC of Ag NPs of average size 8 nm (see Table 1 for details). Similarly, we used the model to reproduce the studies conducted by Pannerselvam et al.[34] and Sharma et al.[35], which evaluated the ZOI of Ag NPs and CuO NPs respectively (see SI, Table S15). The evaluation showed that the predicted values were close to the actual values with the average percentage difference of 14.82%. The experimental result showed selectivity in the MIC and ZOI values of some specific NPs for various bacterial strains. Meanwhile, our platform ML reinforced GA was able to reproduce these patterns and identify NPs with even higher levels of selectivity.

The potential mechanism of action of SANPs may be attributed to a complex interaction between the NPs and cellular components of microbial strains. Several studies have demonstrated that NPs show more efficacy against gram positive bacteria than against the gram negative due to differences in composition of bacterial cell walls[14,36], but there are also studies suggesting the opposite.[11,37] The components of cell wall assist attachment for some NPs while act as protective layers for the others, showing variation in NP toxicity.[7,36] Moreover, various NP properties including their size, shape, type and surface charge interact differently with various cellular components of microorganisms.[38,39] Different strains of bacteria and fungi also have varied defense mechanisms including the ion efflux system, electrostatic repulsion, and biofilms.[40] The efficacy of these defense mechanisms varies among microbes influencing the overall impact of NP toxicity. Moreover, the metabolic activity and the growth rate of microbes might also contribute to the differences in susceptibility. Slowly growing bacteria are more vulnerable to NP-induced stress, while rapidly growing bacteria are more resilient to it due to various stress-response genes expressed.[41,42] Nonetheless, it is noteworthy that the exact mechanism of action of SANPs is still unknown and needs to be investigated. We see a great potential in discovering the underlying mechanisms by experimental analysis and by expanding our model to a larger chemical space, incorporating a wider set of descriptors such as bacterial metabolic-network, metagenomics, and protein interaction patterns.

Our previous work[27] focused on leveraging ML reinforced GA for discovery of selectively cytotoxic NPs for cancer treatment and expanding upon this foundation, we applied a similar methodology to discovery SANPs. This progression emphasizes our commitment to the selective toxicity (SelTox) research domain, where the focus lies on identifying SelTox NPs for precision therapeutics. Looking forward, we envision broadening our research scope within the SelTox domain to identify SelTox NPs that exhibit synergistic interactions with conventional drugs, thereby increasing their clinical efficacy and therapeutic applications. This work represents a significant step forward in researching SelTox, offering

Table 1 Reproduction of experimental results conducted by Pathak et al.[32] and Smekalova et al.[33] using ML reinforced GA. The rows are sorted by the percentage difference in ascending order.

| Bacteria | Method | MC | Predicted | Diff. | % Diff. | Ref. |
|---|---|---|---|---|---|---|
| *P. aeruginosa* | MBEC | 3.10 | 3.09 | 0.01 | 0.32 | [32] |
| *E. coli* | MBEC | 2.80 | 2.79 | 0.01 | 0.36 | [32] |
| *E. faecalis* | MBEC | 3.10 | 3.12 | 0.02 | 0.65 | [32] |
| *C. albicans* | MIC | 2.49 | 2.47 | 0.02 | 0.80 | [32] |
| *S. aureus* | MBEC | 2.80 | 2.76 | 0.04 | 1.43 | [32] |
| *C. albicans* | MBEC | 2.49 | 2.58 | 0.09 | 3.61 | [32] |
| *S. aureus* | MIC | 1.89 | 1.78 | 0.11 | 5.82 | [32] |
| *S. enterica* | MIC | 1.10 | 1.17 | 0.07 | 6.36 | [32] |
| *S. aureus* | MIC | 1.40 | 1.25 | 0.15 | 10.71 | [32] |
| *APP** | MIC | 1.40 | 1.24 | 0.16 | 11.43 | [32] |
| *E. faecalis* | MIC | 1.59 | 1.90 | 0.31 | 19.50 | [33] |
| *P. aeruginosa* | MIC | 1.59 | 1.95 | 0.36 | 22.64 | [33] |
| *E. coli* | MIC | 1.59 | 1.99 | 0.40 | 25.16 | [33] |
| *S. uberis* | MIC | 0.80 | 1.06 | 0.26 | 32.50 | [33] |
| *P. Multocida* | MIC | 2.00 | 1.20 | 0.80 | 40.00 | [33] |
| *E. coli* | MIC | 0.8 | 1.16 | 0.36 | 45.00 | [33] |

Note- MC: minimal concentration, Diff.: difference between actual and predicted MC, % Diff.: percentage difference, Ref.: reference, *APP**: *A. pleuropneumoniae*

a new approach for precision medicine and tailored therapeutic interventions.

In conclusion, this study implemented ML-reinforced GA to discover SANPs, offering potential therapeutic interventions for the infections caused by pathogenic bacteria without disrupting beneficial microbiome communities. We first compiled two datasets, one with the target MC parameter, and the other one with the target ZOI parameter. Then, we screened the best performing CatBoost regression model by evaluating 43 regression models for predicting the antimicrobial activity. Upon hyperparameter optimization, the model achieved 10-fold CV $R^2$ scores of 0.82 and 0.84 on the MC and the ZOI datasets, with the corresponding RMSE values of 0.46 and 2.41. Moreover, we evaluated the model's performance on the test datasets, and the model performed consistently across various NPs and bacterial strains. Afterwards, a GA was developed and optimized to selectively screen the antimicrobial NPs. Leveraging the ML reinforced GA, we identified the selectively antibacterial CuO and ZnO NPs against *Staphylococcus aureus* and *Klebsiella pneumoniae* with the difference in MBC of 392.85 $\mu$g/ml and 285.49 $\mu$g/ml compared to *Bacillus subtilis*. Furthermore, we examined the resilience and the applicability of the ML reinforced GA by using the model to reproduce the experimental results described in literature. These findings contribute significantly to the newly opened research direction of the SelTox NPs. In the future, we aim to expand this research domain to identifying SelTox NPs exhibiting synergistic activity with drugs. This expansion will not only broaden the scope of SelTox NPs but also assist in designing safe and efficient NPs that could be approved and used with drugs for therapeutic applications, such as cancer and microbial infection treatment.

This work was financially supported by the Russian Science Foundation no. 21-73-10150. The authors thank Priority 2030 Federal Academic Leadership Program for infrastructure support. Also, we specially thank Yulia Ryabukhina for additional manuscript curation.

## Author Contributions

Jyakhwo S.: data collection, preprocessing, curation, visualization and analysis, model development and optimization, results interpretation and validation, manuscript writing, graphical design; Bocharova V.: data collection, preprocessing, model development, manuscript writing; Serov N.: study design, manuscript curation, results interpretation; Dmitrenko A.: study design, manuscript writing, results interpretation, supervision; Vinogradov V.: concept development, manuscript curation, supervision, funding.

## Conflicts of interest

There are no conflicts to declare.

## Notes and references

1 S. Gnat, D. Łagowski, A. Nowakiewicz and M. Dyląg, *Journal of Applied Microbiology*, 2021, **131**, 2095–2113.
2 K. Hou, Z. X. Wu, X. Y. Chen, J. Q. Wang, D. Zhang, C. Xiao, D. Zhu, J. B. Koya, L. Wei, J. Li and Z. S. Chen, *Signal Transduction and Targeted Therapy 2022 7:1*, 2022, **7**, 1–28.
3 C. McKernan, T. Benson, S. Farrell and M. Dean, *JAC-Antimicrobial Resistance*, 2021, **3**, year.
4 W. Huang, F. Tao, F. Li, M. Mortimer and L. H. Guo, *NanoImpact*, 2020, **20**, 100268.
5 T. Bruna, F. Maldonado-Bravo, P. Jara and N. Caro, *International Journal of Molecular Sciences*, 2021, **22**, year.
6 S. Shamaila, N. Zafar, S. Riaz, R. Sharif, J. Nazir and S. Naseem, *Nanomaterials*, 2016, **6**, year.
7 M. Ahamed, H. A. Alhadlaq, M. A. Khan, P. Karuppiah and N. A. Al-Dhabi, *Journal of Nanomaterials*, 2014, **2014**, year.
8 A. Sirelkhatim, S. Mahmud, A. Seeni, N. H. M. Kaus, L. C. Ann, S. K. M. Bakhori, H. Hasan and D. Mohamad, *Nano-Micro Letters*, 2015, **7**, 219.
9 T. G. Chatzimitakos and C. D. Stalikas, *Journal of proteome research*, 2016, **15**, 3322–3330.
10 P. Zhang, S. Wu, J. Li, X. Bu, X. Dong, N. Chen, F. Li, J. Zhu, L. Sang, Y. Zeng, S. Liang, Z. Yu and Z. Liu, *Theranostics*, 2022, **12**, 4818.
11 M. Ozdal and S. Gurkok, *ADMET DMPK*, 2022, **10**, 115.
12 W. Gao, S. Thamphiwatana, P. Angsantikul and L. Zhang, *Wiley interdisciplinary reviews. Nanomedicine and nanobiotechnology*, 2014, **6**, 532–547.
13 B. Luan, T. Huynh and R. Zhou, *Nanoscale*, 2016, **8**, 5750–5754.
14 L. Wang, C. Hu and L. Shao, *International Journal of Nanomedicine*, 2017, **12**, 1227.
15 W. Wang, F. Wu, Q. Zhang, N. Zhou, M. Zhang, T. Zheng, Y. Li and B. Z. Tang, *ACS nano*, 2022, **16**, year.
16 N. Lewinski, V. Colvin and R. Drezek, *Small (Weinheim an der Bergstrasse, Germany)*, 2008, **4**, 26–49.
17 U. Kadiyala, N. A. Kotov and J. S. VanEpps, *Current pharmaceutical design*, 2018, **24**, 896.
18 M. Mirzaei, I. Furxhi, F. Murphy and M. Mullins, *Nanomaterials*, 2021, **11**, 1774.
19 G. Mancardi, A. Mikolajczyk, V. K. Annapoorani, A. Bahl, K. Blekos, J. Burk, Y. A. Çetin, K. Chairetakis, S. Dutta, L. Escorihuela, K. Jagiello, A. Singhal, R. van der Pol, M. A. Bañares, N. V. Buchete, M. Calatayud, V. I. Dumit, D. Gardini, N. Jeliazkova, A. Haase, E. Marcoulaki, B. Martorell, T. Puzyn, G. J. A. Sevink, F. C. Simeone, K. Tämm and E. Chiavazzo, *Materials Today*, 2023, **67**, 344–370.
20 E. Goldberg, M. Scheringer, T. D. Bucheli and K. Hungerbühler, *Environmental Science: Nano*, 2015, **2**, 352–360.
21 A. Saadat, A. D. Varniab and S. M. Madani, *Journal of Nanomaterials*, 2022, **2022**, year.
22 L. Zhu, D. W. Pearson, S. L. Benoit, J. Xie, J. Pant, Y. Yang, A. Mondal, H. Handa, J. Y. Howe, Y. C. Hung, J. E. Vidal, R. J. Maier and Y. Zhao, *Nanomaterials*, 2020, **10**, 1–17.
23 N. Zhang, G. Xiong and Z. Liu, *Frontiers in Bioengineering and Biotechnology*, 2022, **10**, year.
24 H. Li, Q. Chen, J. Zhao and K. Urmila, *Scientific Reports*, 2015, **5**, year.
25 W. B. Zhao, K. K. Liu, Y. Wang, F. K. Li, R. Guo, S. Y. Song and C. X. Shan, *Advanced Healthcare Materials*, 2023, **12**, 2300324.
26 F. Vatansever, W. C. de Melo, P. Avci, D. Vecchio, M. Sadasivam, A. Gupta, R. Chandran, M. Karimi, N. A. Parizotto, R. Yin, G. P. Tegos and M. R. Hamblin, *FEMS microbiology reviews*, 2013, **37**, 955.
27 S. Jyakhwo, N. Serov, A. Dmitrenko and V. V. Vinogradov, *Small*, 2023, 2305375.
28 S. J. Belfield, J. W. Firman, S. J. Enoch, J. C. Madden, K. E. Tollefsen and M. T. Cronin, *Computational Toxicology*, 2023, **25**, 100251.
29 A. Banerjee, M. Chatterjee, P. De and K. Roy, *Chemometrics and Intelligent Laboratory Systems*, 2022, **227**, 104613.
30 Y. Li, Z. Wang, L. Tan, L. Liang, S. Liu, J. Huang, J. Lin, K. Peng, Z. Wang, Q. Li, W. Jian, B. Xie, Y. Gao and J. Zheng, *BMC Infectious Diseases*, 2024, **24**, 1–13.
31 O. Ruuskanen, E. Lahti, L. C. Jennings and D. R. Murdoch, *The Lancet*, 2011, **377**, 1264–1275.
32 T. K. Pathak, R. E. Kroon, V. Craciun, M. Popa, M. C. Chifiriuc and H. C. Swart, *Heliyon*, 2019, **5**, year.
33 M. Smekalova, V. Aragon, A. Panacek, R. Prucek, R. Zboril and L. Kvitek, *Veterinary journal (London, England : 1997)*, 2016, **209**, 174–179.
34 B. Pannerselvam, T. S. Alagumuthu, S. K. Cinnaiyan, N. A. Al-Dhabi, K. Ponmurugan, M. Saravanan, S. V. Kanth and K. P. Thangavelu, *Journal of Cluster Science*, 2021, **32**, 63–76.
35 S. Sharma and K. Kumar, *Journal of Dispersion Science and Technology*, 2021, 1–13.
36 A. Sarwar, H. Katas, S. N. Samsudin and N. M. Zin, *PloS one*, 2015, **10**, year.
37 S. Manzoor, D. J. Bashir, K. Imtiyaz, M. M. Rizvi, I. Ahamad, T. Fatma, N. B. Agarwal, I. Arora and M. Samim, *RSC Advances*, 2021, **11**, 24900–24916.
38 F. Aflakian, F. Mirzavi, H. T. Aiyelabegan, A. Soleimani, J. G. Navashenaq, I. Karimi-Sani, A. R. Zomorodi and R. Vakili-Ghartavol, *European Journal of Pharmaceutical Sciences*, 2023, **188**, 106515.
39 A. Ivask, A. Elbadawy, C. Kaweeteerawat, D. Boren, H. Fischer, Z. Ji, C. H. Chang, R. Liu, T. Tolaymat, D. Telesca, J. I. Zink, Y. Cohen, P. A. Holden and H. A. Godwin, *ACS Nano*, 2014, **8**, 374–386.
40 M. F. S. Orozco, N. Niño-Martínez, G. A. Martínez-Castañón, F. T. Méndez and F. Ruiz, *International Journal of Molecular Sciences 2019, Vol. 20, Page 2808*, 2019, **20**, 2808.
41 Y. A. Helmy, K. Taha-Abdelaziz, H. A. E. H. Hawwas, S. Ghosh, S. S. AlKafaas, M. M. Moawad, E. M. Saied, I. I. Kassem and A. M. Mawad, *Antibiotics 2023, Vol. 12, Page 274*, 2023, **12**, 274.
42 S. T. Khan, J. Musarrat and A. A. Al-Khedhairy, *Colloids and Surfaces B: Biointerfaces*, 2016, **146**, 70–83.

ITMO UNIVERSITY
MINISTRY OF SCIENCE AND HIGHER EDUCATION
OF THE RUSSIAN FEDERATION

**ITMO University**

Kronverksky Pr. 49, St. Petersburg,
Russian Federation, 197101
Phone: +7 (812) 232-97-04 | Fax: +7 (812) 232-23-07
international@itmo.ru | en.itmo.ru

_____№_____

To Prof. Dr. Douglas Stephan

Editors-in-Chief of Chemical Communications

Dear Professor: Douglas Stephan

Herewith we submit our manuscript entitled "**SelTox: Discovering the Capacity of Selectively Antimicrobial Nanoparticles for Targeted Eradication of Pathogenic Bacteria**" for consideration as an article to *RSC ChemComm.*

**Motivation:** Microbial infections caused by bacteria, viruses, or fungi pose significant health risks, with antibiotics being the primary treatment. However, their non-selective nature and overuse contribute to antibiotic resistance. Inorganic nanoparticles (NPs) offer a promising alternative due to their ability to combat infections by disrupting bacterial metabolism, damaging cell membranes, and inhibiting biofilm formation. At the same time, size, shape, roughness, surface charge, and other specific properties of NPs facilitate varied interactions with bacteria compared to antibiotics which have a specific mechanism of action. This diversity in mechanisms of action of NPs poses challenges in developing a generalized theory to predict the antimicrobial effects across various bacterial strains. The need becomes more pronounced when considering selectively toxic nanoparticles tailored to exert antimicrobial activity exclusively against pathogenic bacteria. These challenges spur investigations into developing theoretical models based on a data-driven approach, allowing correlation among poorly formalized parameters. Although machine learning has been used for predicting antimicrobial activity of NPs, overall, to our knowledge, no study has evaluated its capability *to discover selectively antimicrobial NPs*.

**Approach:** Our approach involved compiling comprehensive databases to characterize nanoparticles and their antibacterial activity. We implemented machine learning (ML) reinforced genetic algorithm (GA), a screening platform that enables high-throughput identification of NPs with selective antimicrobial activity. This innovative strategy has paved the way for a new research direction focused on identifying NPs possessing selective toxicity (SelTox) properties.

**Results:** In this study, employing ML reinforced GA, we have for the first time identified potential NPs exhibiting selective antimicrobial properties against pathogenic bacteria. The predictive CatBoost regressor model was trained on the unique datasets consisting of 489 samples for minimal concentration prediction and 609 samples for zone of inhibition prediction. The model achieved a mean cross-validation $R^2$ score of 0.82 and 0.84, with RMSE of 0.46 and 2.41 respectively. By combining ML model with GA, we identified

CuO NP (with key parameters including NP synthesis with green methods, nanorods in shape, average NP size of 30 nm, and reaction time of 2 hours) as one of the best selectively antimicrobial NPs. It showed a minimal bactericidal concentration (MBC) of 62.52 µg/ml against *Staphylococcus aureus* whereas it achieved MBC of 455.37 µg/ml for *Bacillus subtilis*. Hence, the selectively antimicrobial CuO NP demonstrated higher toxicity against pathogenic *Staphylococcus aureus* compared to non-pathogenic *Bacillus subtilis*, with a concentration difference of 392.85 µg/ml. The antimicrobial selectivity of NPs is attributed to a complex interaction between the NPs and diverse microbial strains.

**Impact:** The findings of this study present a novel and promising methodology for identifying selectively antimicrobial NPs capable of eradicating pathogenic bacteria without harming non-pathogenic ones. We envision this research will pioneer the discovery of selectively antimicrobial NPs, leading to an expansion of our research focus within the SelTox domain. Furthermore, this expansion will aim to identify selectively toxic NPs capable of synergistically interacting with conventional drugs, thereby enhancing their clinical efficacy and expanding their therapeutic applications. This research marks a substantial advancement in SelTox research, presenting a new approach for precision medicine and tailored therapeutic interventions.

We firmly believe that this research will not only expand the understanding of selectively toxic NPs but also contribute to the development of safe and effective NPs suitable for approval and use in conjunction with drugs for therapeutic purposes, including cancer and microbial infection treatment.

We are confident that this manuscript will capture the interest of the readership of the journal of ***RSC ChemComm***. All authors have thoroughly reviewed the manuscript and have approved its submission. The work presented in the manuscript is original, has not been published previously, and is not under consideration for publication elsewhere.

We propose the following colleagues as potential referees:

Prof. Vadim Kessler (vadim.kessler@slu.se)

Prof. David Avnir (david.avnir@mail.huji.ac.il)

Prof. Natalia L Klyachko (NLKlyachko@enzyme.chem.msu.ru)

Prof. Eugenia Kumacheva (eugenia.kumacheva@utoronto.ca )

Prof. Valentine Ananikov (val@ioc.ac.ru)

Prof. Alexander Kabanov (kabanov@email.unc.edu)

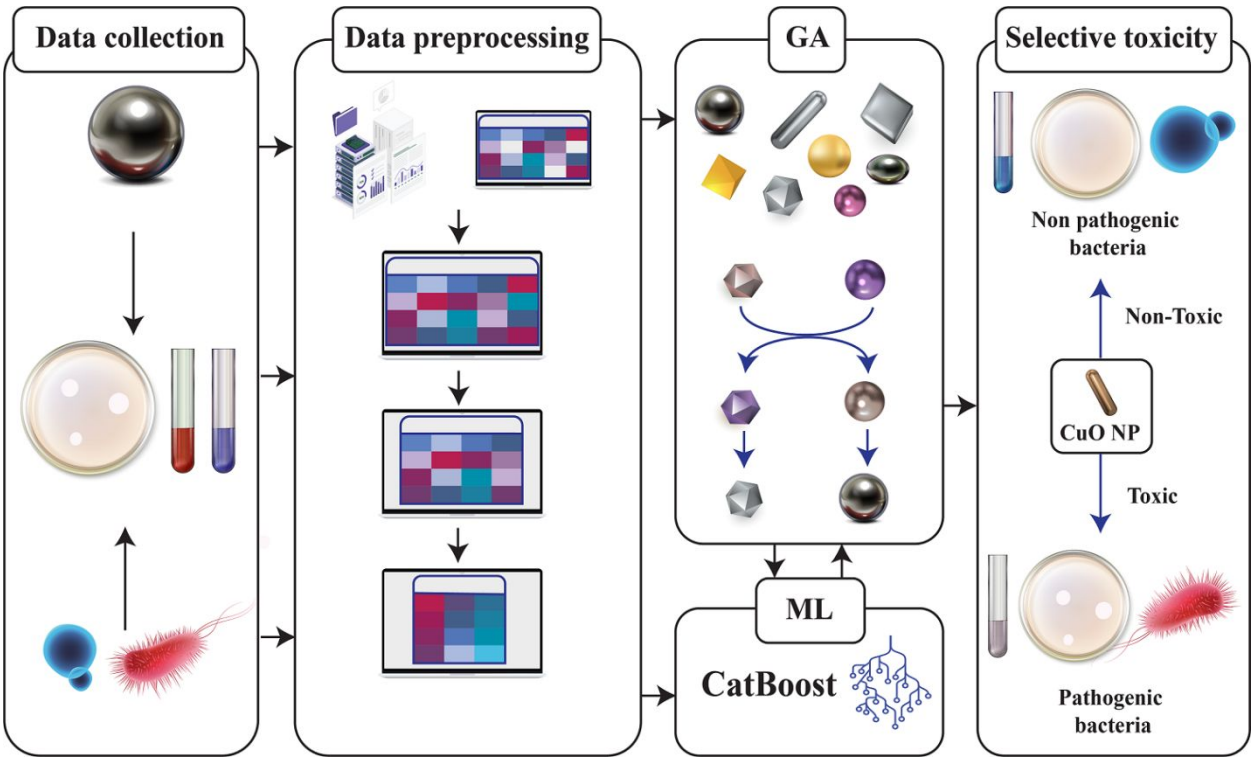
We look forward to hearing from you.

Sincerely,

Prof. Dr. Vladimir V. Vinogradov

ITMO University

**ChemBio cluster**




This study proposes an innovative approach to discover selectively antimicrobial nanoparticles (SANPs) for targeted eradication of pathogenic bacteria while minimizing harm to non-pathogenic ones, using machine learning reinforced genetic algorithm. As a proof of concept, CuO SANPs were identified for the targeted eradication of *Klebsiella pneumoniae*.

Supporting Information

# SelTox: Discovering the Capacity of Selectively Antimicrobial Nanoparticles for Targeted Eradication of Pathogenic Bacteria

*Susan Jyakhwo[1], Valentina Bocharova[1], Nikita Serov[1], Andrei Dmitrenko[1]*, and Vladimir V. Vinogradov[1]***

[1]International Institute "Solution Chemistry of Advanced Materials and Technologies", ITMO University, 191002 Saint Petersburg, Russia

**Corresponding author; Email: vinogradov@scamt-itmo.ru

*Co-corresponding author; Email: dmitrenko@scamt-itmo.ru

## Section A. Description

### 1. Database

### 1.1 Data collection

During the database building, articles related to evaluation of toxicity of metal and metal oxide nanoparticles on different strains of bacteria and fungi were collected using keywords such as "nanoparticles toxicity", "metal and metal oxide nanoparticles", "antibacterial and antimicrobial activity", "biocidal activity", and "synergistic activity" in the period of 2010-2023. Experimental data from more than 70 research articles were collected from different scientific journals such as Nature, MDPI, PubMed, and Scopus while material descriptor and bacterial descriptors were collected using RDKit, NCBI taxonomy, and manual extraction methods. The collected data were separated into two unique datasets: MC dataset and ZOI dataset, for predicting minimal concentration (MC) and zone of inhibition (ZOI). It was essential to use separate databases due to differences in number of features and target parameters, preventing their unification. The MC dataset was created for prediction of minimal bactericidal concentration (MBC), minimal biofilm eradication concentration (MBEC), and minimal inhibitory concentration (MIC) of NP required for eradication or inhibition of bacterial growth. Similarly, the ZOI dataset was created for prediction of the inhibition zone. During this split, data with target variable growth inhibition (GI) %, half maximal inhibitory

concentration (IC50), bacterial viability (%) and others were excluded due to fewer sample size and absence of standardized descriptors for unification.

These datasets comprise the information on NP characteristics such as their types, shape, size, synthesis method, physicochemical properties, and bacterial characteristics including their types, growth conditions, incubation period, and taxonomic descriptors. However, zeta potential, surface charge, coating, bacterial resistance, and other features were omitted as more than 5% of the values were missing.

**1.2 Data preprocessing**

On both datasets standard preprocessing steps were applied. Initially, descriptors with varying units were converted into a single standard unit. Then, columns with similar meaning but different names were combined into one. To address missing values in average NP size and incubation period, gaps were filled by taking the average value of maximum and minimum NP size and incubation period respectively. There were different types of NP synthesis methods - we categorized them into two groups, green synthesis, and chemical synthesis. Similarly, a new category of bacterial type was created and based on the toxic nature of bacteria to the human population, they were classified as pathogenic, non-pathogenic, opportunistic, and allergenic.

In the MC dataset, the concentration ranges from 0-12500 µg/ml were retained, while data with zero concentration were discarded. Afterwards, concentration values were log transformed using base 10 logarithm. Similarly, in the ZOI dataset, samples with inhibition diameter below 6mm were discarded as diameter below 6mm are not experimentally feasible to measure. Furthermore, duplicate entities as well as samples with values missing were removed from the remaining data.

Prior to feature encoding, the distribution of values in each feature was evaluated. In numerical columns, outliers were removed using quantile function. Approximately 1% of data from numerical column concentration (µg/ml), NP size (nm), time set (h), and incubation period (h) were removed as the data distribution was not uniform. The NP concentration of > 10000 µg/ml, size of > 100 nm, time set of > 24 hr, and incubation period of > 48hr were marked as outliers in the MC dataset. For the ZOI dataset, the ZOI of < 6mm, concentration of > 25000 µg/ml, size of > 120 nm, and incubation period of > 252 hr were treated as outliers. Next, we employed the correlation matrix threshold of $|r| > 0.95$ to remove low variance features and highly correlated features, mitigating the risk of overfitting.

For model training, categorical features were encoded using a label encoder to convert them into numerical representations. Further, a standard scaler function was applied to scale numerical columns, ensuring equal impact of each feature by standardizing mean value to 0

with a standard deviation of 1. This normalization reduced the influence of features with larger values, thereby enhancing performance and stability of ML models.

**2. Machine learning**

We adopted a two-step approach for model selection and optimization to identify the most effective model based on $R^2$ score, RMSE, MAE, and MSE. Initially, we trained 43 different ML models with default parameters on both MC and ZOI datasets. The preprocessing steps dramatically increased performance of all models, however, the top three most performing ones were CatBoost Regressor, XGB Regressor, and ExtraTrees Regressor. CatBoost Regressor model had the highest adjusted $R^2$ score 0.7 and the lowest RMSE value 0.49 for the preprocessed dataset for predicting MC. Similarly, the XGB Regressor model had the highest adjusted $R^2 = 0.85$ score and the lowest RMSE = 2.16 score predicting the ZOI. Subsequently, these two top performing models, CatBoost and XGB regressors, were optimized using Bayesian optimization techniques (using Optuna package) to find the best combination of hyperparameters (Table S16 & S17).

Due to non-uniform distribution of parameters, we implemented StratifiedShuffleSplit function to ensure that the train-test split accurately represented the data distribution. Then, the performance of optimized models was evaluated by 10-fold cross validation (for reproducibility, random seed = 42 is used). Both models achieved similar performance with CatBoost regressor exhibiting slightly superior 10-fold CV and lower RMSE score (Table S18 & S19). Thus, we selected the CatBoost regression model for predicting the MC and the ZOI of the NPs.

As we got two datasets, two optimized hyperparameters were used to predict MC and ZOI respectively. For the MC dataset, the target variables were MIC, MBC, and MBEC. Since all of them are minimal concentrations that either inhibit bacterial growth or eradicate bacteria at provided concentration and have the same unit, they were unified and predicted together. Similarly, for the ZOI dataset, the target variable is diameter of inhibition zone that is measured using disk diffusion method and well diffusion method. Unifying MC and ZOI was not feasible because they have different metrics with different independent variables and the attempt to build a single model for prediction will compromise the accuracy of the model. Thus, two separate models were optimized for the specific tasks of predicting minimal concentration and inhibition zone diameter. Notably, target variables such as growth inhibition (%), viability (%), metabolic activity (%), and membrane permeability were excluded. The decision to focus on MC and ZOI was driven by the abundance of experimental data available in literature and the fact that experiments conducted for other target variables often yield results for either MC or

ZOI parameters. Moreover, MC and ZOI are widely recognized parameters in antimicrobial activity testing, offering specific and quantitative measures of NP toxicity against various microorganisms.

In the current model version, descriptors encompassing NP physicochemical properties, experimental conditions, and bacterial properties were incorporated. However, features related to NP interaction, metabolic activity, nanocomposites, and synergistic activities were omitted due to their complexity and the limited availability of standardized descriptors in these domains. Nonetheless, our future effort is directed toward expanding the model with comprehensive descriptors once these descriptors get more standardized.

**3. Genetic algorithm**

Genetic algorithm (GA), a computational approach inspired by the principle of Darwinian evolution and natural selection, are extensively implemented in various fields including ML, to address optimization problems by mimicking the process of natural selection within populations of potential solutions. Here we employed GA for screening and identifying selectively antimicrobial NPs.

Initially, the GA generated a population of NP candidates, each with unique features. In the population, antimicrobial activity of NPs was calculated by optimized CatBoost regressor model that we developed. Subsequently, the fitness of each NPs was calculated by difference in log value of minimal concentration for pathogenic and non-pathogenic bacteria on MC dataset while fitness of NPs was calculated by difference of inhibition zone for ZOI dataset, using following formula,

For MC dataset, Fitness score = $\log_{10}(MC_{np}) - \log_{10}(MC_p)$,

For ZOI dataset, Fitness score = $Z_p - Z_{np}$,

Where $MC_{np}$ is the predicted minimal concentration of non-pathogenic bacteria, $MC_p$ is the predicted minimal concentration of pathogenic bacteria, $Z_p$ is predicted ZOI of pathogenic bacteria and $Z_{np}$ is predicted ZOI of non-pathogenic bacteria. (In the MC dataset, the concentration range spanned from 0 to 12500 µg/ml, with the mean value of 507.5 µg/ml. As the data covers a large range of values, we opted to use logarithms instead of the actual values to bring the range to a more manageable scale.)

Next, unique NP individuals are sorted based on fitness score and half of lower scoring population were discarded emulating the evolutionary concept ‘survival of fittest’ through selective pressure in each generation and gap was filled by adding new individuals keeping population size constant. This process of removing half of individuals reduced the genetic diversity of NPs but at the same time it converged the solution to identify individual with higher

fitness score faster. Moreover, crossover and mutation were introduced in each successive generation for feature exchange and population variation to improve fitness score. The process continued until a specified generation number is reached.

### 3.1 GA optimization

The optimization of GA involved evaluating various parameters such as population size, generation number, mutation, and crossover rate. Change in fitness score was evaluated by varying population size from 10 to 100 in an increment of 10. Similarly, change in fitness score was measured by change in generation number up to 100. Moreover, the effect of mutation rate and crossover rate in fitness score was assessed from 1 to 20%. The optimal parameters for discovering the selectively antimicrobial NPs in the MC dataset were determined to be the population size of 100, the generation number of 60, and the mutation and crossover rate of 1%. For the ZOI dataset, the optimal parameters were the population size of 50, the generation number of 100, and the mutation and crossover rate of 10% (Figure S11 & S12).

As anticipated, population size, and generation numbers exhibited a square root function with fitness score. Fitness increased rapidly at the beginning due to extensive exploration of chemical spaces and subsequent discovery of optimal solutions. However, as generations progressed, genetic diversity diminished, posing challenges in identifying significantly improved solutions. Further increment in the population size beyond these points plateaued fitness score improvement which might be because increase in population size caused algorithm to encounter similar solutions and no further improvement was possible. Moreover, the reduction in mean fitness score with increase in population size indicated that only handful of individuals possess selective antimicrobial properties. Overall, higher values for GA parameters (population size, generation number, mutation, and crossover rate) resulted in minimal increase in fitness score while lower led to suboptimal results.

### 3.2 GA candidate generation

The developed GA is capable of screening over 500 samples per second on a personal computer equipped with i7 11800H CPU and 16GB RAM. Utilizing the optimal parameters of GA, we identified top selectively antimicrobial NPs that could be utilized to eradicate pneumonia causing pathogenic bacteria *Staphylococcus aureus* and *Klebsiella pneumoniae*. We identified various unique NPs such as CuO, ZnO, $TiO_2$, and Ag that have selective antimicrobial activity against pathogenic bacteria with minimal effect on non-pathogenic bacteria. (see details on TableS11, S12, S13, and S14)

**4. Python library used:**

| | | | |
|---|---|---|---|
| Pycharm | 2023.2.5 | optuna | 3.5.0 |
| catboost | 1.2.2 | pandas | 2.2.0 |
| category-encoders | 2.6.3 | scikit-learn | 1.4.0 |
| joblib | 1.3.2 | scipy | 1.12.0 |
| lazypredict | 0.2.12 | seaborn | 0.13.2 |
| matplotlib | 3.8.2 | shap | 0.44.1 |
| numpy | 1.26.3 | xgboost | 2.0.3 |

## Section B. Supplementary figures

a) b)

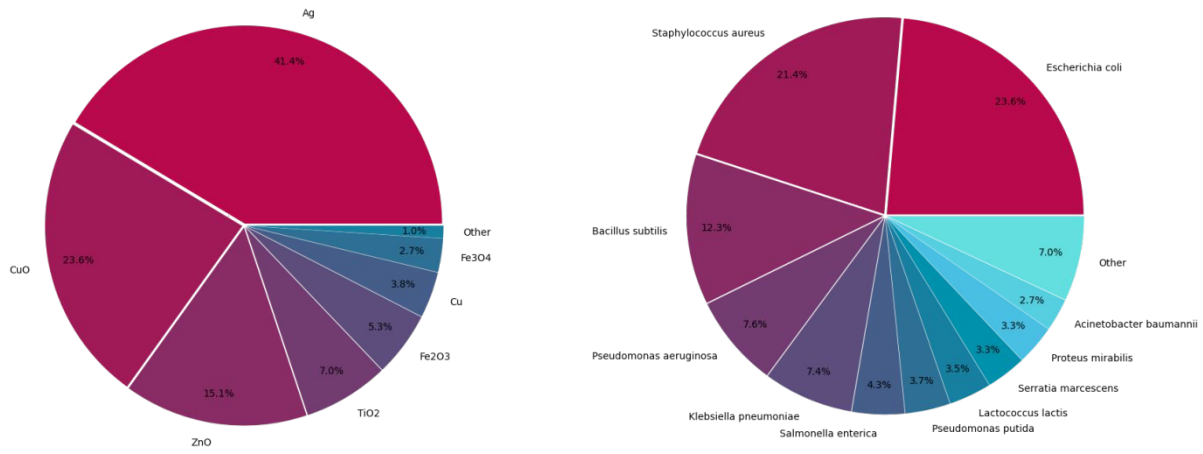



Figure S1. Data distribution of a) NPs and b) bacteria in ZOI dataset.

a)

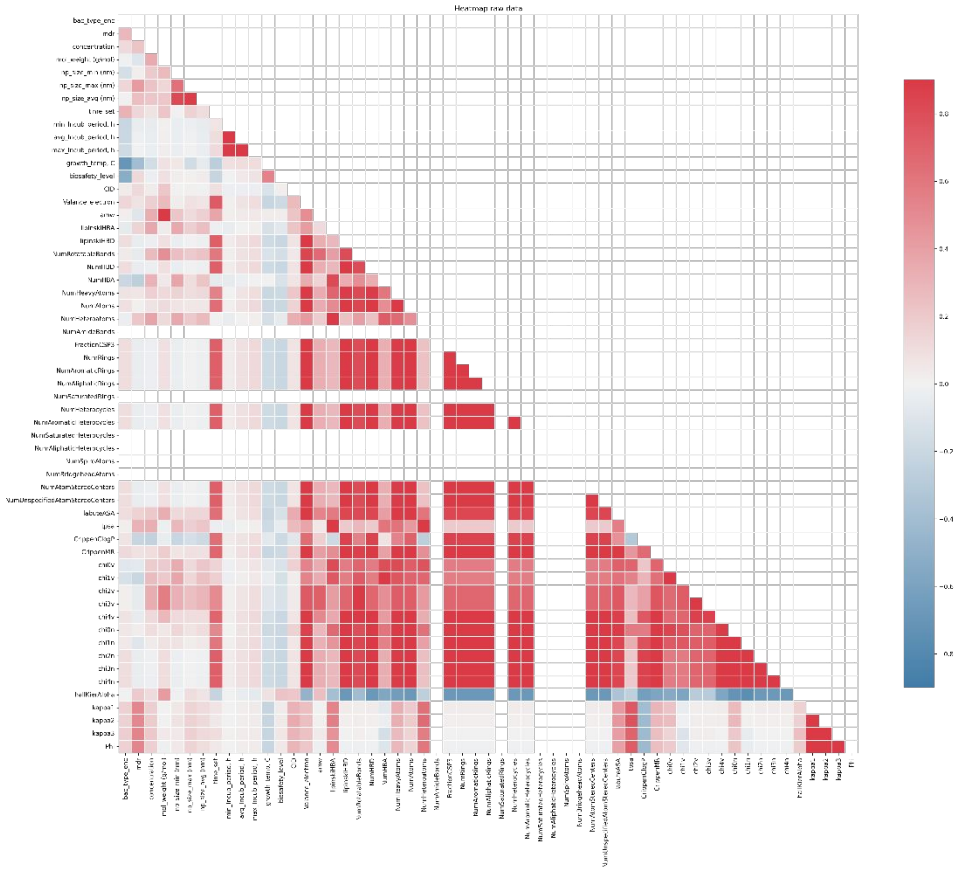



Figure S2. a) Correlation matrix of raw data colored by Pearson's coefficient in MC dataset.

b)

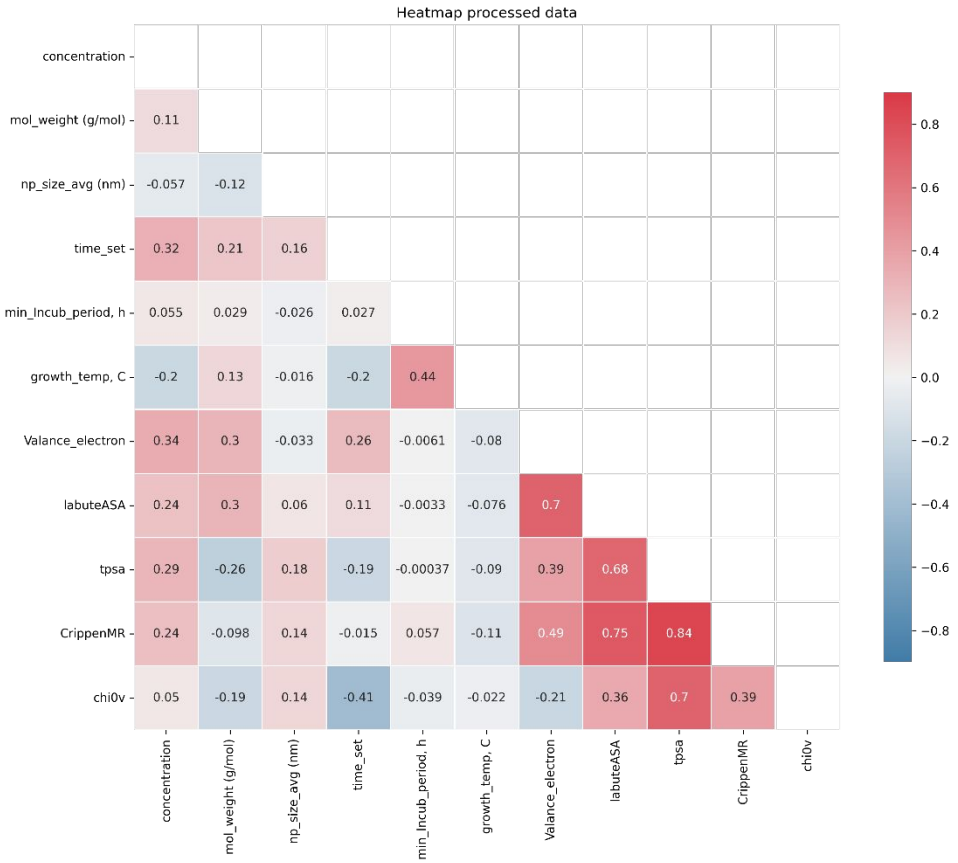
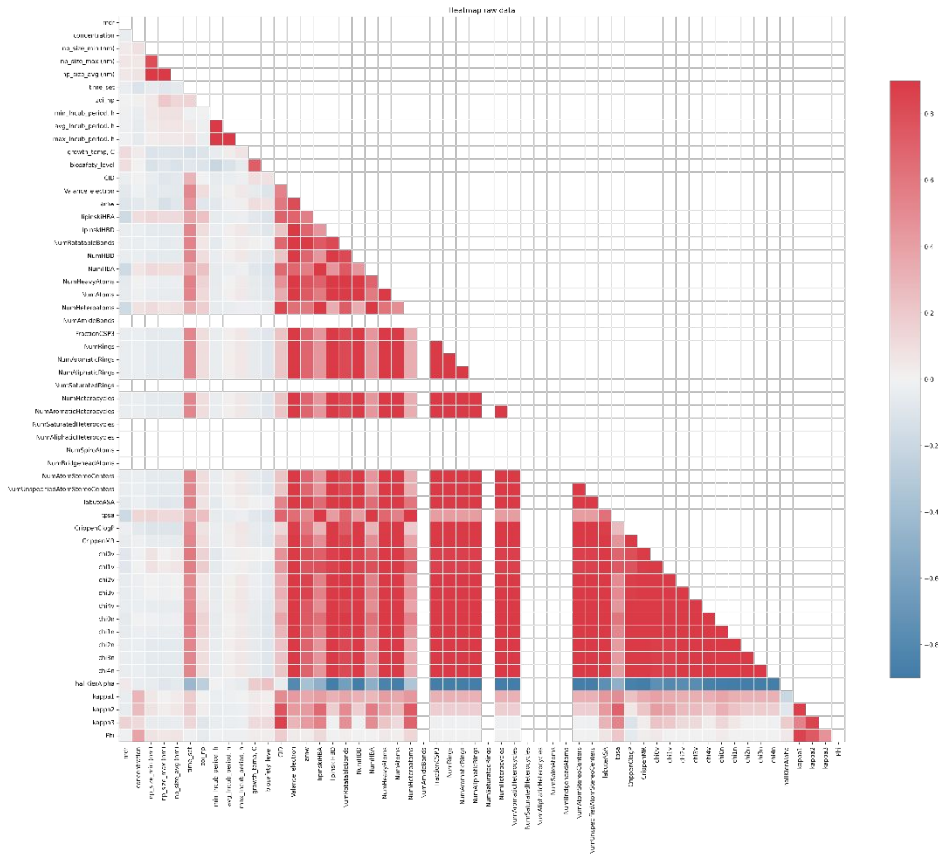



Figure S2. b) Correlation matrix of preprocessed data colored by Pearson's coefficient in MC dataset.

a)



b)

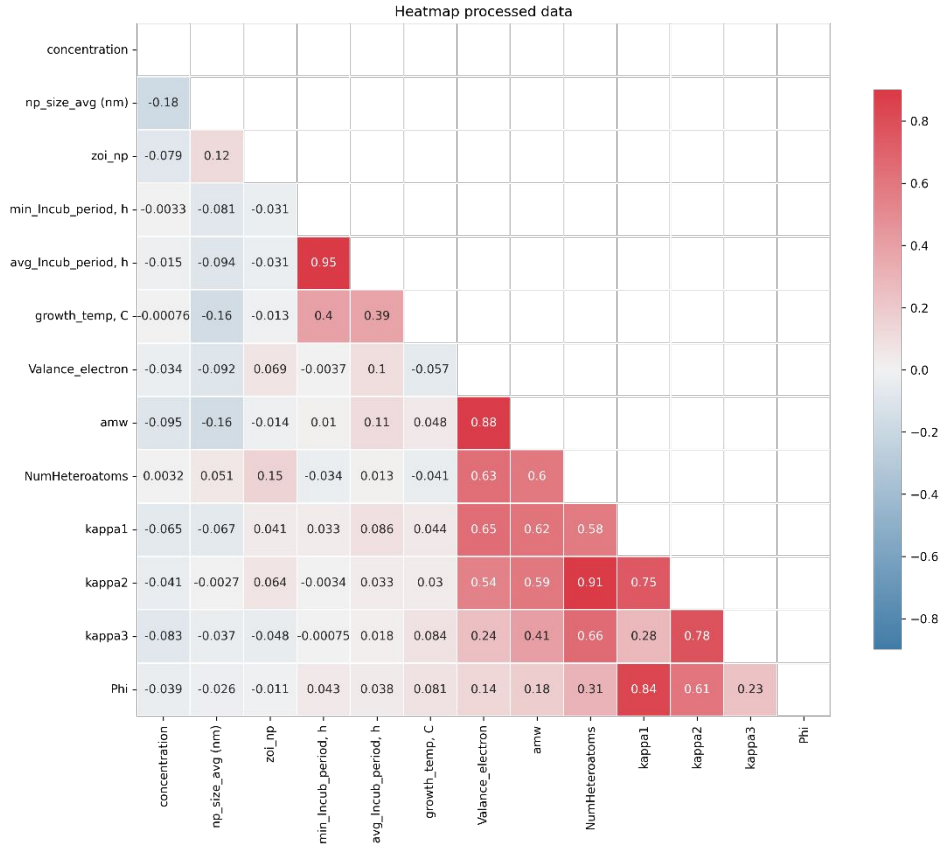



Figure S3. Correlation matrix of a) raw data and b) preprocessed data colored by Pearson's coefficient in ZOI dataset.

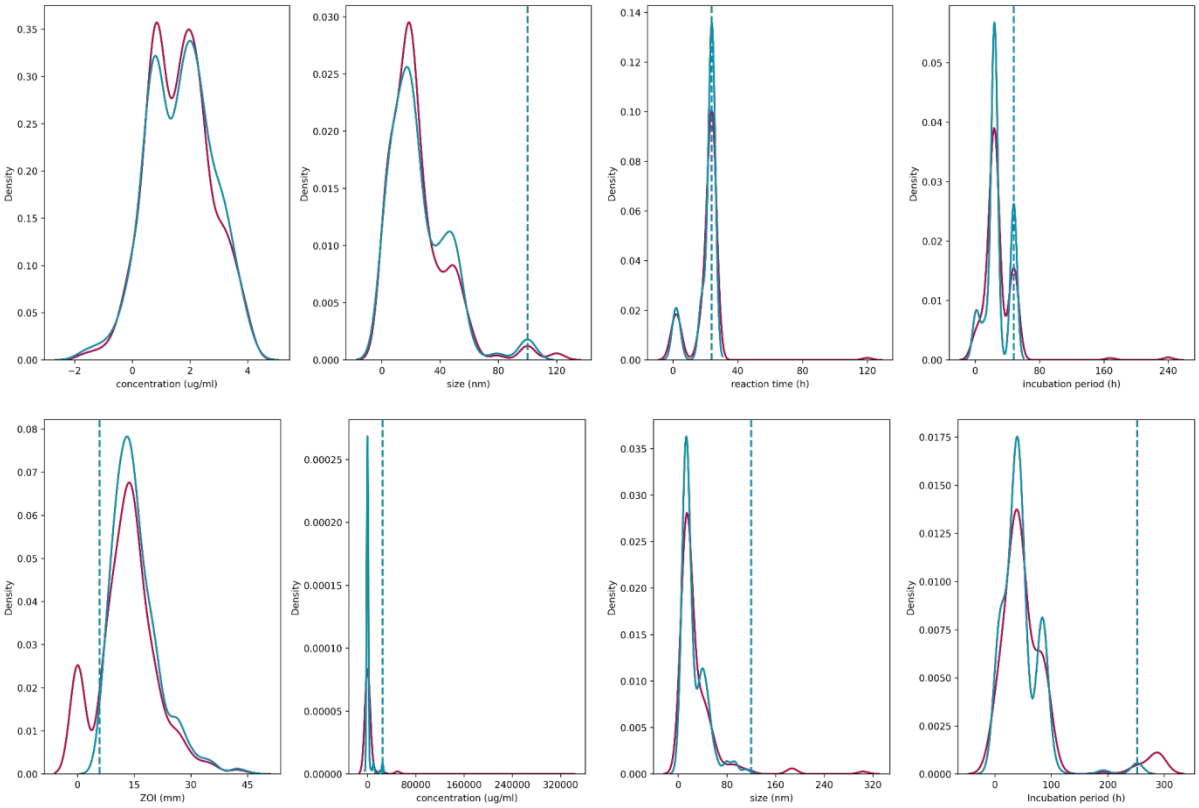



Figure S4. Kde plot of a) concentration, size, time set and incubation period of raw and preprocessed data (MC dataset) and b) ZOI, concentration, size and incubation period of raw and preprocessed data (ZOI dataset).

a)

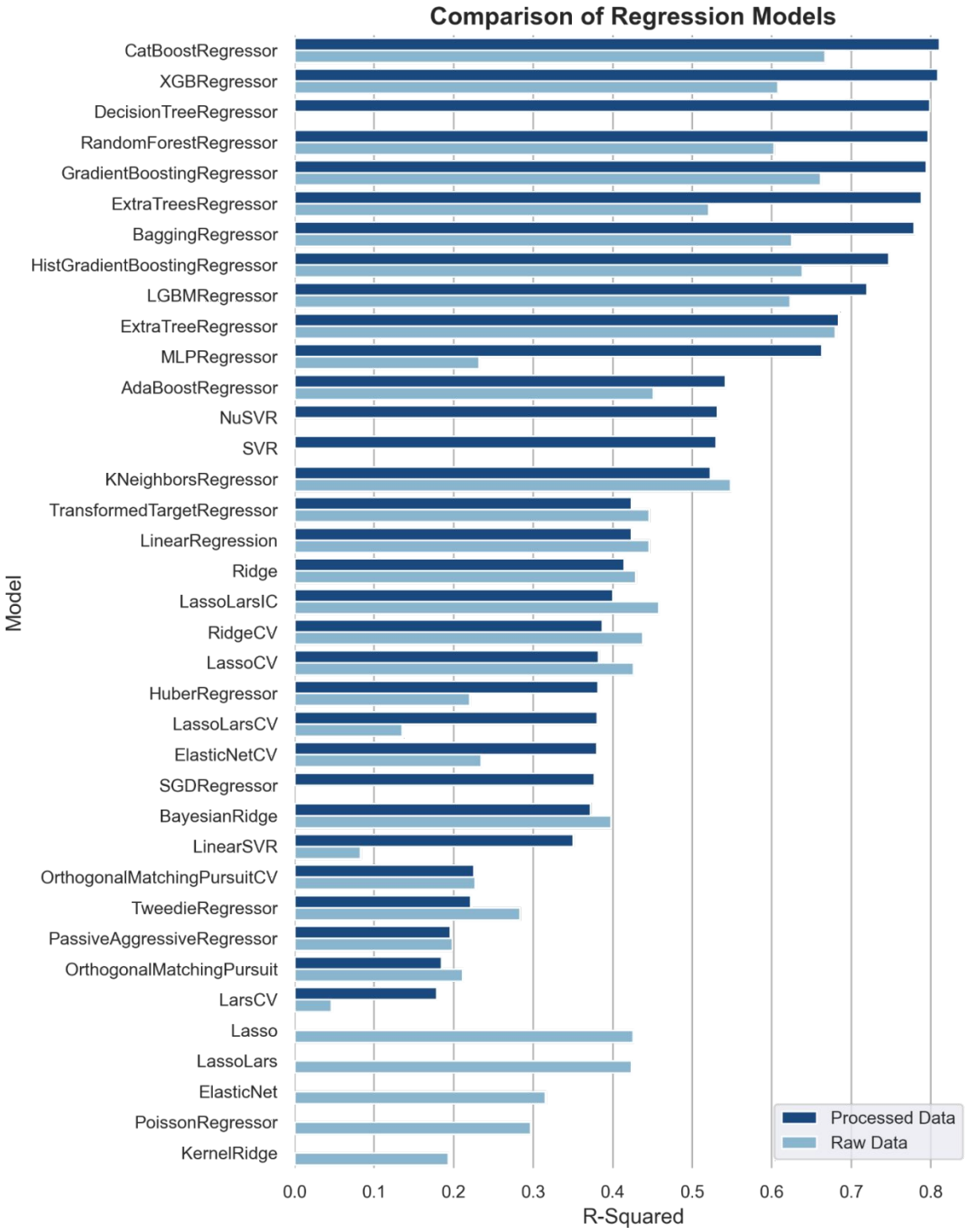



Figure S5. a) Comparison of R2 score of 43 regression models on raw and preprocessed MC dataset.

b)

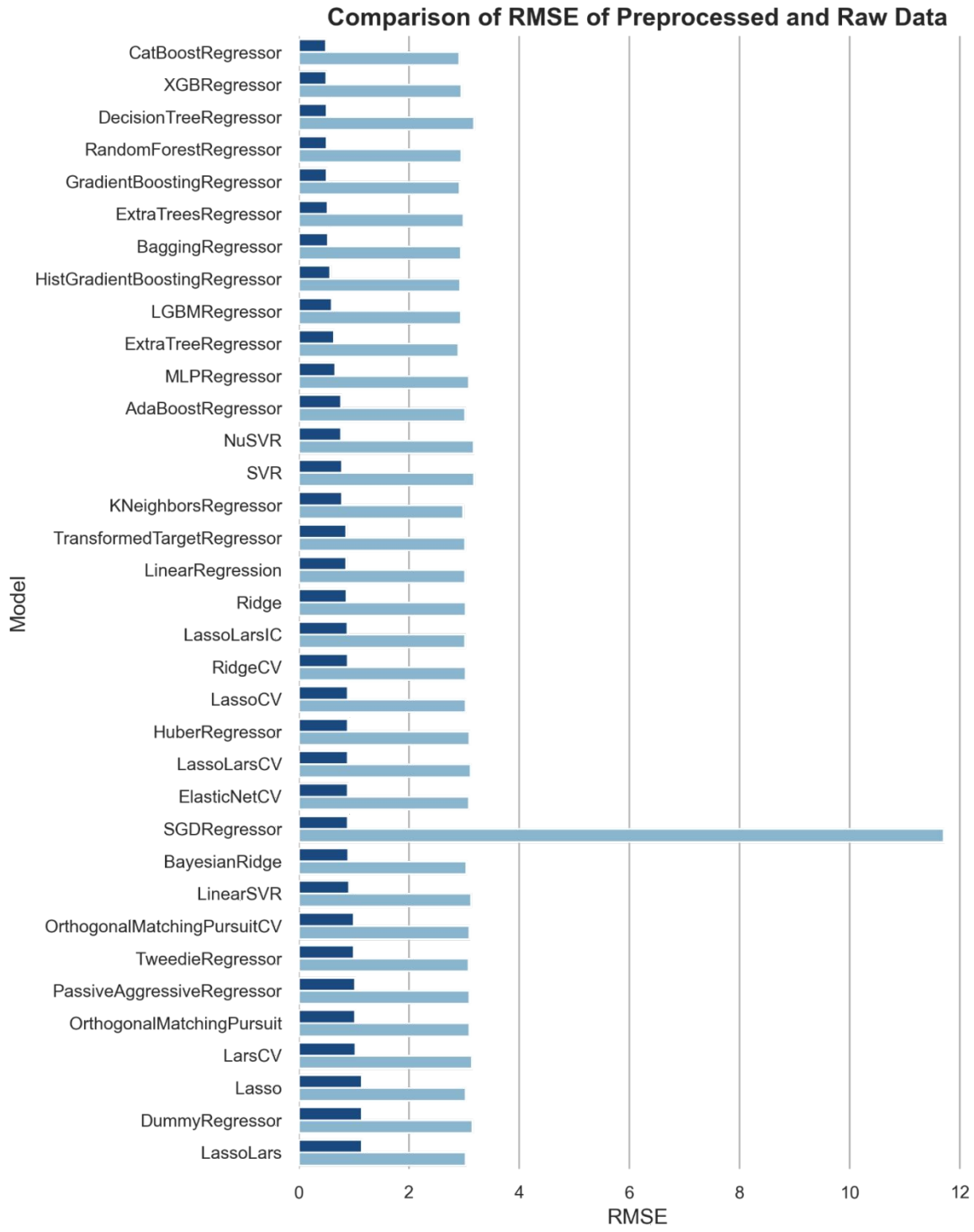



Figure S5. b) Comparison of RMSE score of 43 regression model on raw and preprocessed MC dataset.

a)

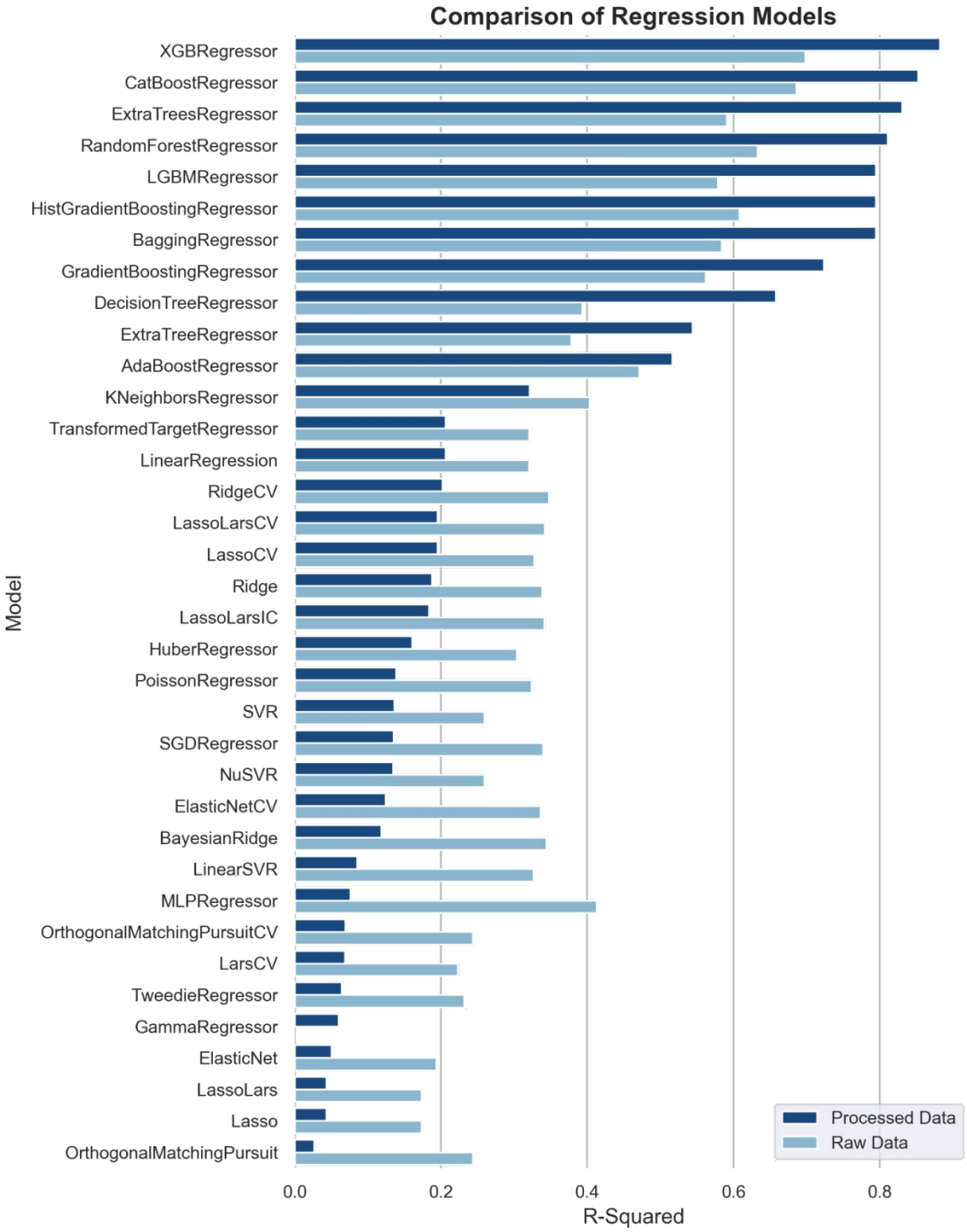



Figure S6. a) Comparison of R2 score of 43 regression model on raw and preprocessed ZOI dataset.

b)

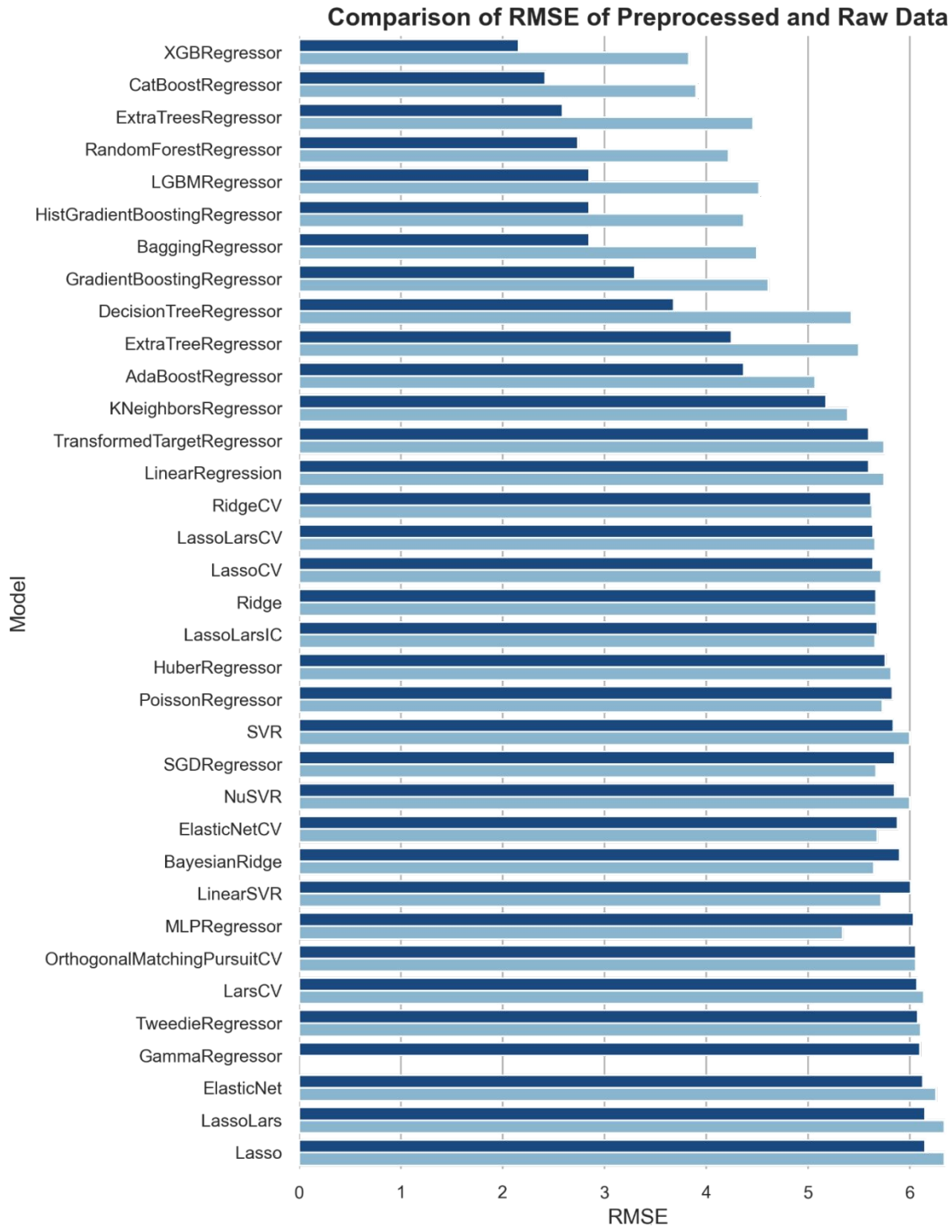



Figure S6. b) Comparison of RMSE score of 43 regression model on raw and preprocessed ZOI dataset.

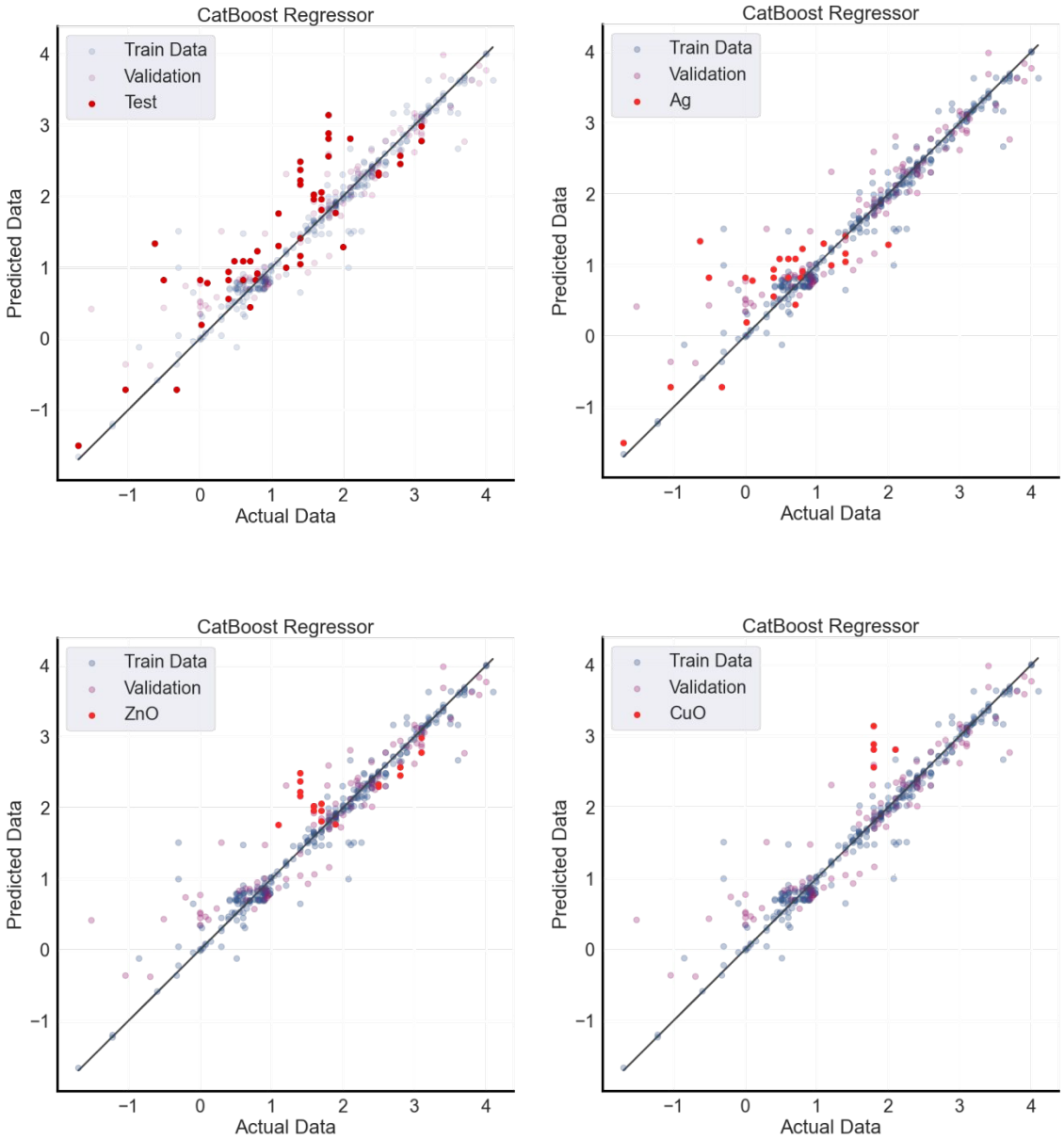



Figure S7. Prediction of MC values for test data and specific NPs (Ag, ZnO and CuO).

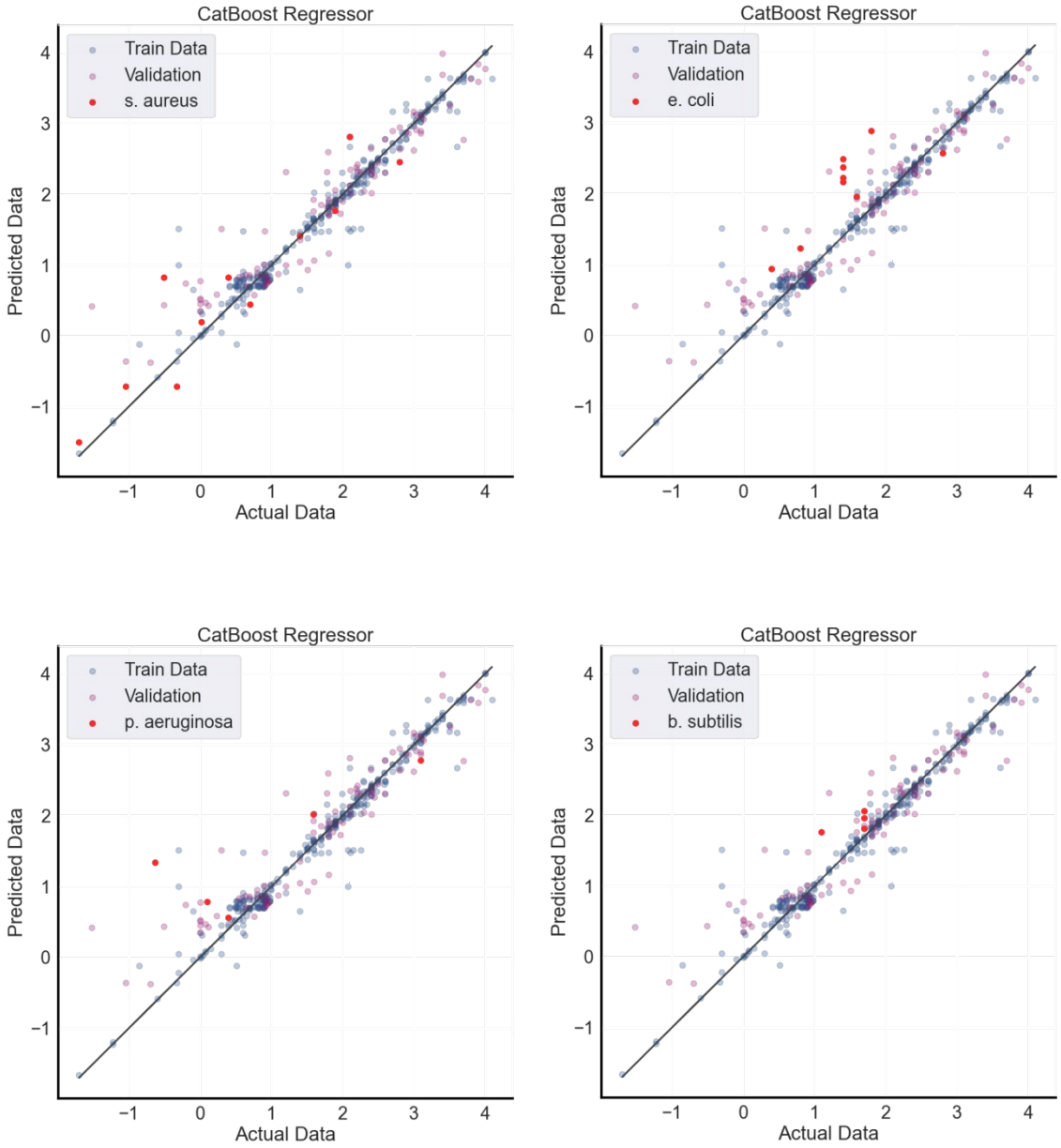



Figure S8. Prediction of MC values for various bacterial strains (*Staphylococcus aureus, Escherichia coli, Pseudomonas aeruginosa, Bacillus subtilis*).

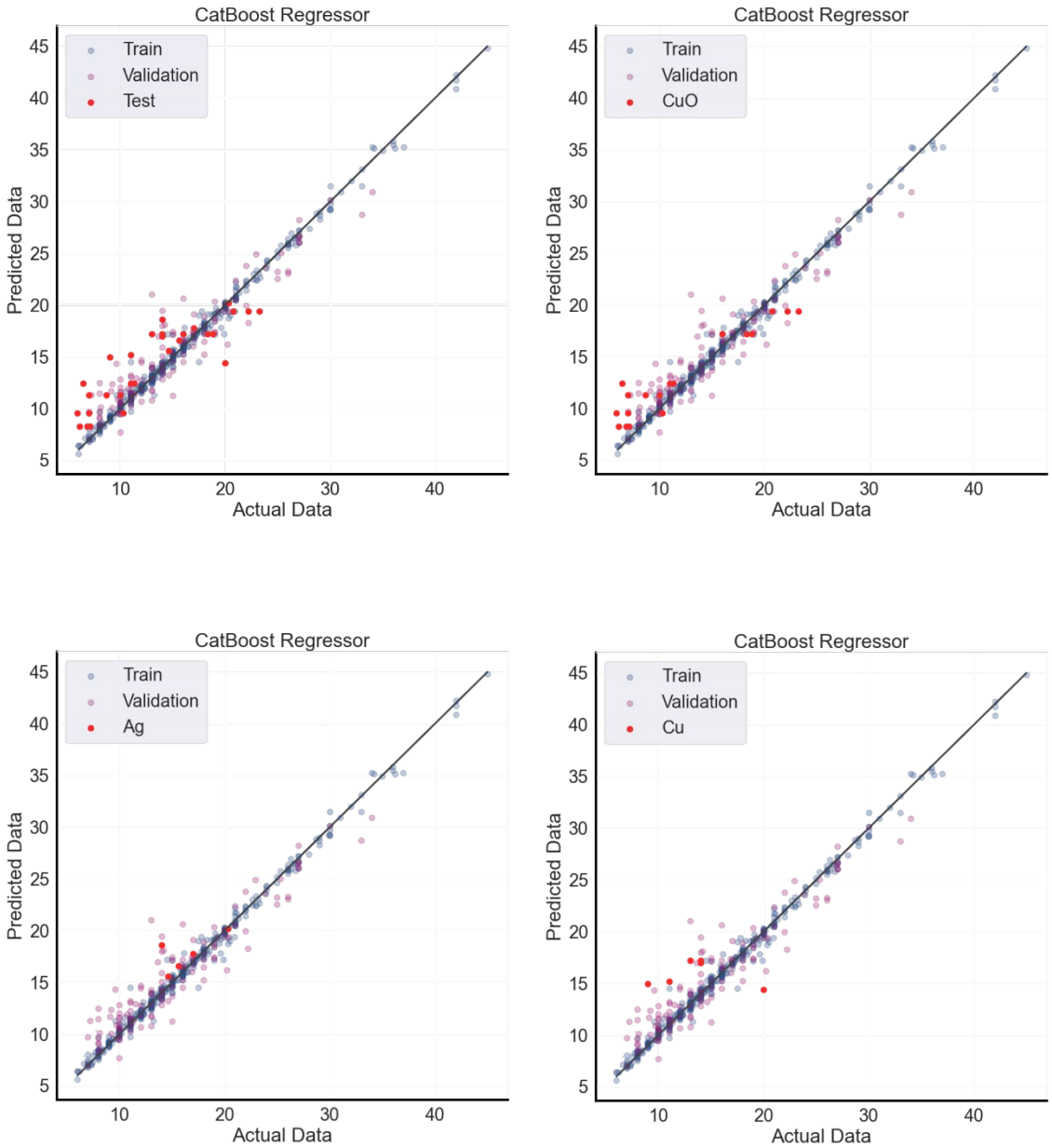



Figure S9. Prediction of ZOI values for test data and specific NPs (CuO, Ag, and Cu).

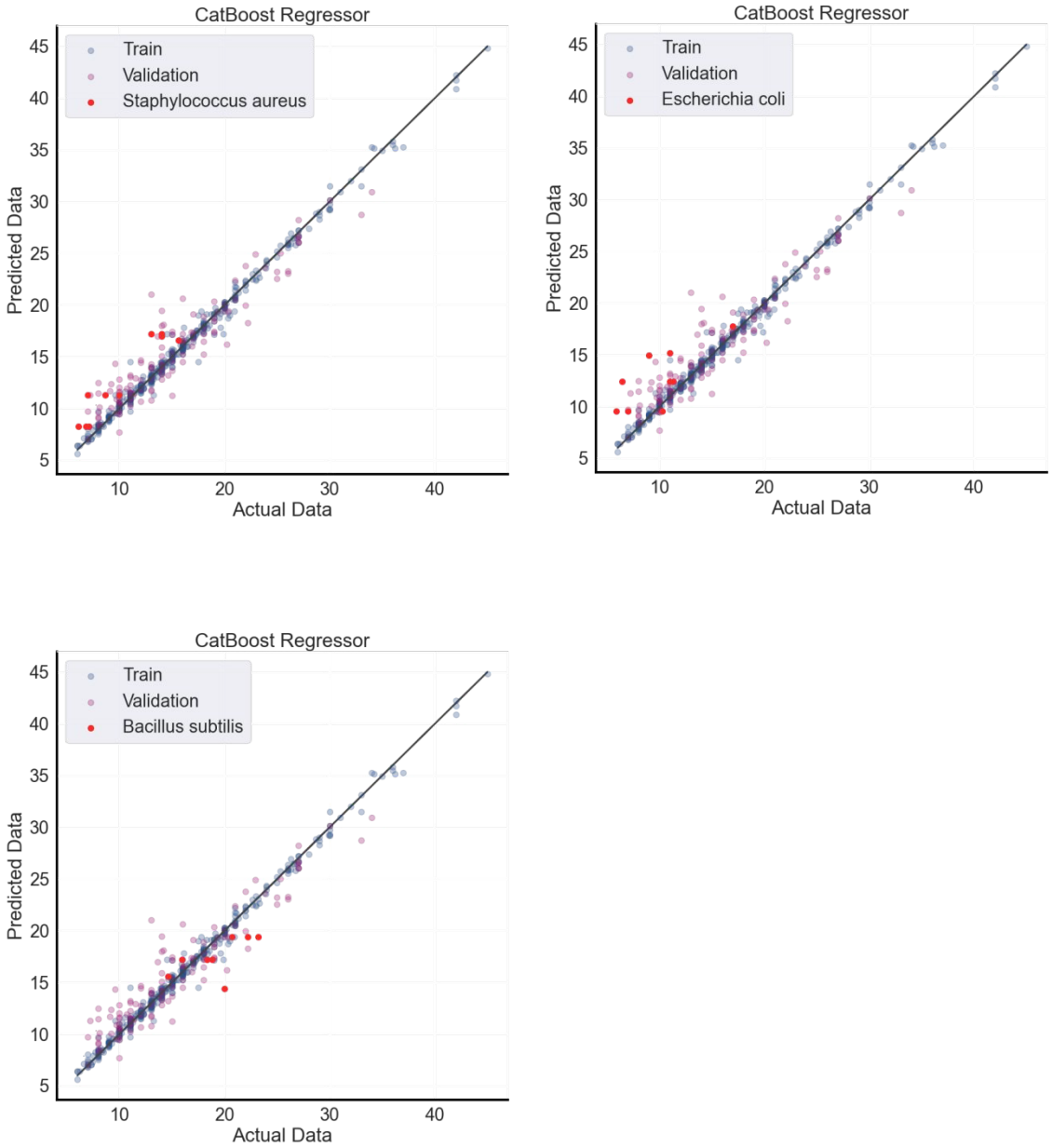



Figure S10. Prediction of ZOI values for various bacterial strains (*Staphylococcus aureus, Escherichia coli, Bacillus subtilis*).

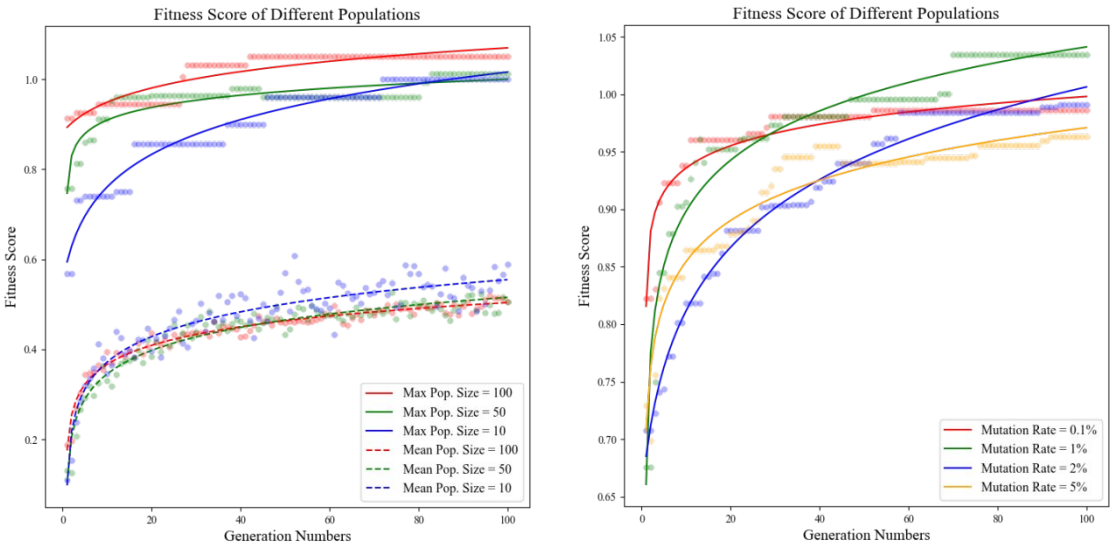



Figure S11. GA optimization by changing a) generation number and population size b) mutation and crossover rate for MC dataset.

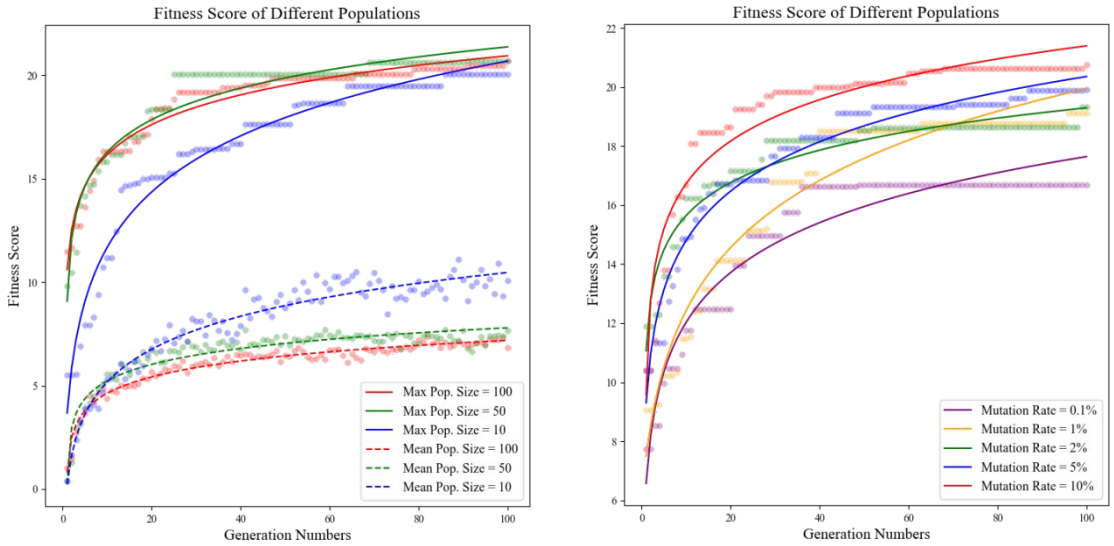



Figure S12. GA optimization by changing b) generation number and population size b) mutation and crossover rate for ZOI dataset.

**Section C. Supplementary tables**

Table S1. List of Independent variables.

| **Experimental conditions** | **Material descriptors** | **Bacterial descriptors** |
|---|---|---|
| 1. Concentration (µg/ml)<br>2. Average NP size (nm)<br>3. NP synthesis method<br>4. Time set (hr)<br>5. Zone of inhibition (mm) | 1. NP type<br>2. Shape<br>3. Mol_weight<br>4. Valence electron<br>5. LabuteASA<br>6. Tpsa<br>7. CrippenMR<br>8. Chi0v<br>9. NumHeteroatoms<br>10. Kappa1<br>11. Kappa2<br>12. Kappa3<br>13. Phi | 1. Bacteria strains<br>2. Bac_type<br>3. Kingdom<br>4. Phylum<br>5. Class<br>6. Order<br>7. Family<br>8. Genus<br>9. Gram<br>10. Isolated_from<br>11. Avg_incub_period<br>12. Growth_temp |

Table S2. Distribution of Categorical features.

| **Categorical features** | **Unique values in (MC dataset)** | **Unique values in (ZOI Dataset)** |
|---|---|---|
| NP type | 13 | 13 |
| Bacteria | 34 | 45 |
| Bacteria type | 3 | 4 |
| np_synthesis | 5 | 3 |
| method | 3 | 3 |
| shape | 10 | 15 |
| kingdom | 2 | 2 |
| phylum | 4 | 3 |
| class | 6 | 7 |
| order | 11 | 13 |
| family | 16 | 18 |
| genus | 21 | 25 |
| gram | 3 | 3 |
| isolated_from | 13 | 14 |

Table S3. Distribution of numerical features in MC dataset.

| **Numerical features** | **Mean value** | **Standard deviation** | **Min value** | **Max value** |
|---|---|---|---|---|
| log_concentration | 1.66 | 1.11 | -1.70 | 4.10 |
| mol_weight (g/mol) | 107.32 | 37.99 | 58.69 | 231.74 |
| np_size_avg (nm) | 25.79 | 20.43 | 0.80 | 100.00 |
| time_set (hr) | 20.27 | 7.44 | 0.00 | 24.00 |
| min_Incub_period (hr) | 27.63 | 14.03 | 0.16 | 48.00 |
| growth_temp ($^0$C) | 35.73 | 2.70 | 30.00 | 37.00 |
| Valance_electron | 10.92 | 3.46 | 8.00 | 28.00 |
| labuteASA | 23.71 | 5.06 | 17.42 | 51.56 |
| tpsa | 9.99 | 14.07 | 0.00 | 85.50 |
| CrippenMR | 0.57 | 1.73 | 0.00 | 13.57 |
| chi0v | 2.76 | 1.12 | 1.24 | 4.69 |

Table S4. Distribution of numerical features in ZOI dataset.

| **Numerical features** | **Mean value** | **Standard deviation** | **Min value** | **Max value** |
|---|---|---|---|---|
| Concentration (μg/ml) | 1420.51 | 4822.34 | 0.31 | 25000.00 |
| np_size_avg (nm) | 25.39 | 20.33 | 3.00 | 120.00 |
| zoi_np (mm) | 15.49 | 6.31 | 5.89 | 45.00 |
| min_Incub_period (hr) | 25.96 | 23.26 | 0.00 | 168.00 |
| avg_Incub_period (hr) | 47.31 | 37.42 | 0.00 | 252.00 |
| growth_temp ($^0$C) | 34.39 | 3.85 | 22.00 | 37.00 |
| valance_electron | 18.48 | 22.27 | 6.00 | 110.00 |
| amw | 108.09 | 54.93 | 58.93 | 465.96 |
| NumHeteroatoms | 1.86 | 1.19 | 1.00 | 7.00 |
| kappa1 | 3.42 | 4.83 | 1.51 | 36.04 |
| kappa2 | 1.59 | 2.92 | 0.09 | 15.78 |
| kappa3 | 2.41 | 2.94 | 0.79 | 19.65 |
| Phi | 3.63 | 16.69 | 0.16 | 144.26 |

Table S5. Model performance on raw data in MC dataset.

| Model | Adjusted R-Squared | R-Squared | RMSE | Time Taken |
|---|---|---|---|---|
| ExtraTreeRegressor | 0.36 | 0.68 | 2.90 | 0.01 |
| CatBoostRegressor | 0.33 | 0.67 | 2.91 | 1.78 |
| GradientBoostingRegressor | 0.32 | 0.66 | 2.92 | 0.16 |
| HistGradientBoostingRegressor | 0.27 | 0.64 | 2.93 | 0.69 |
| BaggingRegressor | 0.25 | 0.63 | 2.94 | 0.05 |
| LGBMRegressor | 0.24 | 0.62 | 2.94 | 0.09 |
| XGBRegressor | 0.21 | 0.61 | 2.95 | 1.35 |
| RandomForestRegressor | 0.20 | 0.60 | 2.95 | 0.45 |
| KNeighborsRegressor | 0.09 | 0.55 | 2.98 | 0.05 |
| ExtraTreesRegressor | 0.04 | 0.52 | 2.99 | 0.29 |
| LassoLarsIC | -0.09 | 0.46 | 3.02 | 0.07 |
| AdaBoostRegressor | -0.10 | 0.45 | 3.02 | 0.04 |
| TransformedTargetRegressor | -0.12 | 0.45 | 3.02 | 0.03 |
| LinearRegression | -0.12 | 0.45 | 3.02 | 0.02 |
| RidgeCV | -0.13 | 0.44 | 3.03 | 0.04 |
| Ridge | -0.15 | 0.43 | 3.03 | 0.02 |
| LassoCV | -0.16 | 0.43 | 3.03 | 0.46 |
| Lasso | -0.16 | 0.43 | 3.03 | 0.04 |
| LassoLars | -0.16 | 0.42 | 3.03 | 0.02 |
| BayesianRidge | -0.21 | 0.40 | 3.04 | 0.08 |
| ElasticNet | -0.38 | 0.32 | 3.07 | 0.06 |
| PoissonRegressor | -0.42 | 0.30 | 3.08 | 0.04 |
| TweedieRegressor | -0.44 | 0.28 | 3.08 | 0.02 |
| ElasticNetCV | -0.54 | 0.23 | 3.09 | 0.08 |
| MLPRegressor | -0.55 | 0.23 | 3.09 | 2.11 |
| OrthogonalMatchingPursuitCV | -0.56 | 0.23 | 3.10 | 0.05 |
| HuberRegressor | -0.57 | 0.22 | 3.10 | 0.06 |
| OrthogonalMatchingPursuit | -0.59 | 0.21 | 3.10 | 0.01 |

| | | | | |
|---|---|---|---|---|
| PassiveAggressiveRegressor | -0.61 | 0.20 | 3.10 | 0.01 |
| KernelRidge | -0.62 | 0.19 | 3.11 | 0.02 |
| LassoLarsCV | -0.74 | 0.14 | 3.12 | 0.04 |
| LinearSVR | -0.85 | 0.08 | 3.13 | 0.01 |
| LarsCV | -0.92 | 0.05 | 3.14 | 0.05 |
| DummyRegressor | -1.01 | 0.00 | 3.15 | 0.01 |
| NuSVR | -1.18 | -0.08 | 3.17 | 0.03 |
| SVR | -1.25 | -0.12 | 3.18 | 0.04 |
| QuantileRegressor | -1.26 | -0.12 | 3.18 | 11.84 |
| DecisionTreeRegressor | -1.30 | -0.14 | 3.18 | 0.01 |
| GaussianProcessRegressor | -12.50 | -5.71 | 3.57 | 0.10 |
| SGDRegressor | -2.57E+17 | -1.27E+17 | 11.71 | 0.01 |
| RANSACRegressor | -1.07E+24 | -5.31E+23 | 15.01 | 0.21 |
| Lars | -2.64E+43 | -1.31E+43 | 24.71 | 0.015 |

Table S6. Model performance on preprocessed data in MC dataset.

| **Model** | **Adjusted R-Squared** | **R-Squared** | **RMSE** | **Time Taken** |
|---|---|---|---|---|
| CatBoostRegressor | 0.75 | 0.81 | 0.49 | 0.77 |
| XGBRegressor | 0.75 | 0.81 | 0.50 | 1.32 |
| DecisionTreeRegressor | 0.73 | 0.80 | 0.51 | 0.01 |
| RandomForestRegressor | 0.73 | 0.80 | 0.51 | 0.20 |
| GradientBoostingRegressor | 0.73 | 0.79 | 0.51 | 0.06 |
| ExtraTreesRegressor | 0.72 | 0.79 | 0.52 | 0.14 |
| BaggingRegressor | 0.71 | 0.78 | 0.53 | 0.02 |
| HistGradientBoostingRegressor | 0.66 | 0.75 | 0.57 | 0.25 |
| LGBMRegressor | 0.63 | 0.72 | 0.60 | 0.04 |
| ExtraTreeRegressor | 0.58 | 0.68 | 0.64 | 0.01 |
| MLPRegressor | 0.55 | 0.66 | 0.66 | 0.30 |
| AdaBoostRegressor | 0.39 | 0.54 | 0.77 | 0.03 |

| | | | | |
|---|---|---|---|---|
| NuSVR | 0.38 | 0.53 | 0.77 | 0.01 |
| SVR | 0.38 | 0.53 | 0.78 | 0.01 |
| KNeighborsRegressor | 0.37 | 0.52 | 0.78 | 0.05 |
| TransformedTargetRegressor | 0.23 | 0.42 | 0.86 | 0.01 |
| LinearRegression | 0.23 | 0.42 | 0.86 | 0.01 |
| Ridge | 0.22 | 0.41 | 0.87 | 0.01 |
| LassoLarsIC | 0.20 | 0.40 | 0.88 | 0.01 |
| RidgeCV | 0.19 | 0.39 | 0.89 | 0.01 |
| LassoCV | 0.18 | 0.38 | 0.89 | 0.07 |
| HuberRegressor | 0.18 | 0.38 | 0.89 | 0.04 |
| LassoLarsCV | 0.18 | 0.38 | 0.89 | 0.02 |
| ElasticNetCV | 0.18 | 0.38 | 0.89 | 0.06 |
| SGDRegressor | 0.17 | 0.38 | 0.89 | 0.01 |
| BayesianRidge | 0.17 | 0.37 | 0.90 | 0.02 |
| LinearSVR | 0.14 | 0.35 | 0.91 | 0.01 |
| OrthogonalMatchingPursuitCV | -0.03 | 0.23 | 1.00 | 0.01 |
| TweedieRegressor | -0.03 | 0.22 | 1.00 | 0.02 |
| PassiveAggressiveRegressor | -0.07 | 0.20 | 1.02 | 0.01 |
| OrthogonalMatchingPursuit | -0.08 | 0.18 | 1.02 | 0.01 |
| LarsCV | -0.09 | 0.18 | 1.03 | 0.02 |
| Lasso | -0.34 | -0.01 | 1.14 | 0.01 |
| DummyRegressor | -0.34 | -0.01 | 1.14 | 0.00 |
| LassoLars | -0.34 | -0.01 | 1.14 | 0.01 |
| ElasticNet | -0.34 | -0.01 | 1.14 | 0.01 |
| QuantileRegressor | -0.38 | -0.04 | 1.15 | 3.30 |
| KernelRidge | -2.12 | -1.35 | 1.73 | 0.01 |
| Lars | -40.15 | -29.96 | 6.30 | 0.01 |
| GaussianProcessRegressor | -4759.48 | -3581.63 | 67.769 | 0.06 |
| RANSACRegressor | -4.51E+24 | -3.39E+24 | 2.09E+12 | 0.15 |

Table S7. Model performance on raw data in ZOI dataset.

| Model | Adjusted R-Squared | R-Squared | RMSE | Time Taken |
|---|---|---|---|---|
| XGBRegressor | 0.44 | 0.70 | 3.83 | 1.68 |
| CatBoostRegressor | 0.42 | 0.69 | 3.90 | 1.95 |
| RandomForestRegressor | 0.32 | 0.63 | 4.22 | 2.55 |
| HistGradientBoostingRegressor | 0.27 | 0.61 | 4.37 | 1.18 |
| ExtraTreesRegressor | 0.24 | 0.59 | 4.46 | 0.43 |
| BaggingRegressor | 0.22 | 0.58 | 4.50 | 0.08 |
| LGBMRegressor | 0.21 | 0.58 | 4.52 | 0.27 |
| GradientBoostingRegressor | 0.18 | 0.56 | 4.61 | 0.24 |
| AdaBoostRegressor | 0.01 | 0.47 | 5.07 | 0.15 |
| MLPRegressor | -0.09 | 0.41 | 5.34 | 3.07 |
| KNeighborsRegressor | -0.11 | 0.40 | 5.39 | 0.09 |
| DecisionTreeRegressor | -0.13 | 0.39 | 5.43 | 0.02 |
| ExtraTreeRegressor | -0.16 | 0.38 | 5.50 | 0.02 |
| RidgeCV | -0.22 | 0.35 | 5.63 | 0.11 |
| BayesianRidge | -0.22 | 0.34 | 5.65 | 0.05 |
| LassoLarsCV | -0.23 | 0.34 | 5.66 | 0.13 |
| LassoLarsIC | -0.23 | 0.34 | 5.66 | 0.06 |
| SGDRegressor | -0.23 | 0.34 | 5.67 | 0.06 |
| Ridge | -0.23 | 0.34 | 5.67 | 0.04 |
| ElasticNetCV | -0.24 | 0.34 | 5.68 | 0.36 |
| LassoCV | -0.25 | 0.33 | 5.72 | 0.52 |
| LinearSVR | -0.26 | 0.33 | 5.72 | 0.03 |
| PoissonRegressor | -0.26 | 0.32 | 5.73 | 0.04 |
| TransformedTargetRegressor | -0.27 | 0.32 | 5.75 | 0.08 |
| LinearRegression | -0.27 | 0.32 | 5.75 | 0.04 |
| HuberRegressor | -0.30 | 0.30 | 5.82 | 0.11 |
| NuSVR | -0.38 | 0.26 | 6.00 | 0.04 |
| SVR | -0.38 | 0.26 | 6.00 | 0.20 |

| | | | | |
|---|---|---|---|---|
| OrthogonalMatchingPursuitCV | -0.41 | 0.24 | 6.06 | 0.04 |
| OrthogonalMatchingPursuit | -0.41 | 0.24 | 6.06 | 0.01 |
| TweedieRegressor | -0.43 | 0.23 | 6.11 | 0.06 |
| LarsCV | -0.45 | 0.22 | 6.14 | 0.15 |
| ElasticNet | -0.50 | 0.19 | 6.26 | 0.01 |
| Lasso | -0.54 | 0.17 | 6.34 | 0.05 |
| LassoLars | -0.54 | 0.17 | 6.34 | 0.01 |
| QuantileRegressor | -0.93 | -0.04 | 7.10 | 23.57 |
| DummyRegressor | -0.94 | -0.04 | 7.11 | 0.01 |
| PassiveAggressiveRegressor | -3.08 | -1.19 | 10.32 | 0.01 |
| KernelRidge | -7.06 | -3.32 | 14.50 | 0.04 |
| GaussianProcessRegressor | -23179.94 | -12427.34 | 777.27 | 0.14 |
| RANSACRegressor | -2.76E+21 | -1.48E+21 | 2.68E+11 | 0.89 |
| Lars | -6.23E+36 | -3.34E+36 | 1.27E+19 | 0.02 |

Table S8. Model performance on preprocessed data in ZOI dataset.

| **Model** | **Adjusted R-Squared** | **R-Squared** | **RMSE** | **Time Taken** |
|---|---|---|---|---|
| XGBRegressor | 0.85 | 0.88 | 2.16 | 1.30 |
| CatBoostRegressor | 0.81 | 0.85 | 2.42 | 0.84 |
| ExtraTreesRegressor | 0.78 | 0.83 | 2.59 | 0.17 |
| RandomForestRegressor | 0.76 | 0.81 | 2.74 | 0.24 |
| LGBMRegressor | 0.74 | 0.79 | 2.85 | 0.04 |
| HistGradientBoostingRegressor | 0.74 | 0.79 | 2.85 | 0.26 |
| BaggingRegressor | 0.74 | 0.79 | 2.85 | 0.03 |
| GradientBoostingRegressor | 0.64 | 0.72 | 3.30 | 0.10 |
| DecisionTreeRegressor | 0.56 | 0.66 | 3.68 | 0.01 |
| ExtraTreeRegressor | 0.41 | 0.54 | 4.25 | 0.01 |
| AdaBoostRegressor | 0.38 | 0.52 | 4.37 | 0.07 |
| KNeighborsRegressor | 0.13 | 0.32 | 5.18 | 0.05 |
| TransformedTargetRegressor | -0.02 | 0.21 | 5.60 | 0.01 |

| | | | | |
|---|---|---|---|---|
| LinearRegression | -0.02 | 0.21 | 5.60 | 0.01 |
| RidgeCV | -0.03 | 0.20 | 5.62 | 0.02 |
| LassoLarsCV | -0.04 | 0.20 | 5.64 | 0.03 |
| LassoCV | -0.04 | 0.20 | 5.64 | 0.14 |
| Ridge | -0.05 | 0.19 | 5.67 | 0.01 |
| LassoLarsIC | -0.05 | 0.18 | 5.68 | 0.02 |
| HuberRegressor | -0.08 | 0.16 | 5.76 | 0.03 |
| PoissonRegressor | -0.11 | 0.14 | 5.83 | 0.02 |
| SVR | -0.11 | 0.14 | 5.84 | 0.02 |
| SGDRegressor | -0.11 | 0.14 | 5.85 | 0.01 |
| NuSVR | -0.11 | 0.13 | 5.85 | 0.01 |
| ElasticNetCV | -0.13 | 0.12 | 5.88 | 0.08 |
| BayesianRidge | -0.13 | 0.12 | 5.90 | 0.03 |
| LinearSVR | -0.18 | 0.09 | 6.01 | 0.01 |
| MLPRegressor | -0.19 | 0.08 | 6.04 | 0.36 |
| OrthogonalMatchingPursuitCV | -0.20 | 0.07 | 6.06 | 0.02 |
| LarsCV | -0.20 | 0.07 | 6.07 | 0.02 |
| TweedieRegressor | -0.20 | 0.06 | 6.08 | 0.01 |
| GammaRegressor | -0.21 | 0.06 | 6.10 | 0.01 |
| ElasticNet | -0.22 | 0.05 | 6.13 | 0.01 |
| LassoLars | -0.23 | 0.04 | 6.15 | 0.01 |
| Lasso | -0.23 | 0.04 | 6.15 | 0.01 |
| OrthogonalMatchingPursuit | -0.25 | 0.03 | 6.20 | 0.01 |
| DummyRegressor | -0.29 | 0.00 | 6.29 | 0.01 |
| QuantileRegressor | -0.35 | -0.05 | 6.43 | 3.82 |
| PassiveAggressiveRegressor | -0.76 | -0.37 | 7.35 | 0.01 |
| KernelRidge | -7.50 | -5.60 | 16.15 | 0.01 |
| GaussianProcessRegressor | -921.55 | -715.69 | 168.32 | 0.03 |
| Lars | -1353379.41 | -1051385.43 | 6446.70 | 0.01 |
| RANSACRegressor | -8.39E+17 | -6.52E+17 | 5076836702 | 0.15 |

Table S9. Metric score evaluation for the test data (MC dataset).

| Test data for MC dataset | Evaluation Metrics | |
|---|---|---|
| | **RMSE** | **MAE** |
| **NP** | | |
| Ag | 0.62 | 0.45 |
| ZnO | 0.52 | 0.43 |
| CuO | 1.02 | 0.99 |
| **Bacteria** | | |
| *Staphylococcus aureus* | 0.52 | 0.39 |
| *Escherichia coli* | 0.76 | 0.70 |
| *Pseudomonas aeruginosa* | 0.97 | 0.967 |
| *Bacillus subtilis* | 0.40 | 0.35 |
| *Staphylococcus epidermidis* | 0.50 | 0.43 |

Table S10. Metric score evaluation for the test data (ZOI dataset).

| ZOI | Evaluation Metrics | |
|---|---|---|
| | **RMSE** | **MAE** |
| **NP** | | |
| CuO | 2.64 | 2.25 |
| Ag | 2.19 | 1.49 |
| Cu | 4.51 | 4.51 |
| **Bacteria** | | |
| *Staphylococcus aureus* | 2.71 | 2.45 |
| *Escherichia coli* | 3.55 | 2.93 |
| *Bacillus subtilis* | 2.79 | 2.32 |

Table S11. Top predicted selectively antimicrobial NPs with main parameters and fitness score on *Bacillus subtilis* Vs *Staphylococcus aureus* (MC dataset). The rows are sorted by the fitness score.

| NP | np_synthesis | method | shape | np_size (nm) | time_set | predicted log_MC | | Fitness | predicted MC (converted) | | difference in MC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | non-pathogenic | pathogenic | | non-pathogenic | pathogenic | |
| ZnO | chemical_synthesis | MBC | rod-shaped | 30.00 | 3.00 | 1.10 | -0.29 | 1.39 | 12.62 | 0.51 | 12.11 |
| ZnO | chemical_synthesis | MBC | spheroidal | 30.00 | 18.00 | 1.46 | 0.09 | 1.36 | 28.64 | 1.24 | 27.40 |
| Pt | chemical_synthesis | MBC | spherical | 30.00 | 3.00 | 1.30 | 0.39 | 0.92 | 20.16 | 2.43 | 17.73 |
| **CuO** | **green_synthesis** | **MBC** | **rod-shaped** | **30.00** | **2.00** | **2.66** | **1.80** | **0.86** | **455.37** | **62.52** | **392.85** |
| ZnO | chemical_synthesis | MBEC | hexagonal | 30.00 | 1.00 | 2.05 | 1.19 | 0.86 | 111.06 | 15.48 | 95.59 |
| Pd | chemical_synthesis | MBC | spherical | 30.00 | 4.00 | 1.32 | 0.49 | 0.83 | 20.77 | 3.10 | 17.67 |
| ZnO | chemical_synthesis | MBEC | rod-shaped | 10.40 | 3.00 | 1.91 | 1.13 | 0.78 | 81.78 | 13.57 | 68.21 |
| Ag | green_synthesis | MBC | spheroidal | 30.00 | 0.00 | 1.30 | 0.53 | 0.77 | 19.92 | 3.38 | 16.54 |
| ZnO | green_synthesis | MBEC | spherical | 8.70 | 3.00 | 2.47 | 1.71 | 0.76 | 293.99 | 50.79 | 243.20 |
| Ag | chemical_synthesis | MBEC | spheroidal | 30.00 | 3.00 | 2.22 | 1.48 | 0.75 | 167.81 | 29.86 | 137.95 |

Table S12. Top predicted selectively antimicrobial NPs with its main parameters and fitness score on *Bacillus subtilis* Vs *Klebsiella pneumoniae* (MC dataset). The rows are sorted by the fitness score.

| NP | np_synthesis | method | shape | np_size (nm) | time_set | predicted log_MC | | Fitness | predicted MC (converted) | | difference in MC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | non-pathogenic | pathogenic | | non-pathogenic | pathogenic | |
| ZnO | green_synthesis | MBC | quasi-hexagonal | 6.00 | 1.00 | 1.36 | 0.49 | 0.87 | 22.89 | 3.09 | 19.79 |
| TiO2 | green_synthesis | MBC | quasi-hexagonal | 6.00 | 4.00 | 2.46 | 1.59 | 0.86 | 285.37 | 39.12 | 246.26 |
| ZnO | green_synthesis | MBEC | quasi-hexagonal | 6.00 | 4.00 | 1.48 | 0.64 | 0.84 | 30.13 | 4.38 | 25.74 |
| Ag | chemical_synthesis | MIC | spheroidal | 6.00 | 18.00 | 1.29 | 0.46 | 0.83 | 19.44 | 2.90 | 16.55 |
| **ZnO** | **green_synthesis** | **MBC** | **cubic** | **6.00** | **4.00** | **2.53** | **1.72** | **0.81** | **337.98** | **52.49** | **285.49** |
| ZnO | green_synthesis | MBEC | quasi-spherical | 5.10 | 4.00 | 2.35 | 1.58 | 0.77 | 226.39 | 38.06 | 188.33 |
| Ag | chemical_synthesis | MIC | spheroidal | 8.70 | 18.00 | 1.23 | 0.50 | 0.73 | 16.85 | 3.14 | 13.71 |
| ZnO | green_synthesis | MBC | quasi-hexagonal | 10.00 | 1.00 | 1.28 | 0.55 | 0.73 | 19.08 | 3.57 | 15.51 |
| TiO2 | green_synthesis | MBC | quasi-hexagonal | 8.00 | 0.00 | 2.48 | 1.76 | 0.72 | 303.47 | 57.46 | 246.01 |
| TiO2 | green_synthesis | MBC | quasi-hexagonal | 8.00 | 3.00 | 2.45 | 1.73 | 0.72 | 281.35 | 54.08 | 227.26 |

Table S 13.Top predicted selectively antimicrobial NPs with its main parameters and fitness score on *Bacillus subtilis* Vs *Staphylococcus aureus* (ZOI dataset). The rows are sorted by the fitness score.

| NP | np_synthesis | method | shape | concentration | np_size (nm) | predicted ZOI | | Fitness |
|---|---|---|---|---|---|---|---|---|
| | | | | | | non pathogenic | pathogenic | |
| **TiO2** | **green_synthesis** | **disk_diffusion** | **hexagonal** | **40.00** | **35.00** | **18.34** | **40.89** | **22.56** |
| CuO | green_synthesis | disk_diffusion | hexagonal | 40.00 | 37.10 | 17.65 | 38.75 | 21.10 |
| ZnO | green_synthesis | disk_diffusion | cubic | 40.00 | 33.00 | 15.44 | 33.95 | 18.51 |
| Co | green_synthesis | disk_diffusion | ellipsoidal | 32.00 | 27.43 | 12.71 | 29.47 | 16.75 |
| Ag | chemical_synthesis | disk_diffusion | ellipsoidal | 40.00 | 37.86 | 15.72 | 31.84 | 16.12 |
| TiO2 | green_synthesis | disk_diffusion | cylindrical | 42.50 | 41.23 | 19.75 | 35.35 | 15.59 |
| TiO2 | green_synthesis | disk_diffusion | Spherical | 50.00 | 15.00 | 11.24 | 25.85 | 14.61 |
| Fe3O4 | green_synthesis | disk_diffusion | ellipsoidal | 32.00 | 41.23 | 16.85 | 31.33 | 14.48 |
| Cu2O | green_synthesis | disk_diffusion | hexagonal | 40.69 | 41.23 | 19.11 | 32.38 | 13.26 |
| Cu2O | green_synthesis | disk_diffusion | hexagonal | 40.69 | 25.90 | 14.50 | 27.53 | 13.03 |

Table S 14. Top predicted selectively antimicrobial NPs with its main parameters and fitness score on Bacillus subtilis Vs Klebsiella pneumoniae (ZOI dataset). The rows are sorted by the fitness score.

| NP | np_synthesis | method | shape | concentration | np_size (nm) | predicted ZOI | | Fitness |
|---|---|---|---|---|---|---|---|---|
| | | | | | | non pathogenic | pathogenic | |
| **Ag** | **chemical_synthesis** | **disk_diffusion** | **spherical** | **30.00** | **10.00** | **19.35** | **29.19** | **9.84** |
| CuO | chemical_synthesis | disk_diffusion | spherical | 30.00 | 10.00 | 19.88 | 29.52 | 9.64 |
| Cu2O | green_synthesis | disk_diffusion | ellipsoidal | 40.00 | 37.10 | 17.40 | 26.34 | 8.94 |
| Ag | chemical_synthesis | well_diffusion | snowflake | 28.00 | 6.00 | 16.98 | 25.24 | 8.27 |
| ZnO | green_synthesis | disk_diffusion | nanorods | 40.69 | 35.00 | 20.00 | 27.66 | 7.66 |
| Ag | green_synthesis | disk_diffusion | spherical | 28.00 | 6.00 | 15.47 | 22.86 | 7.38 |
| TiO2 | green_synthesis | disc_diffusion | hexagonal | 40.00 | 37.86 | 17.78 | 24.83 | 7.05 |
| Ag | chemical_synthesis | well_diffusion | snowflake | 32.00 | 10.00 | 16.08 | 22.40 | 6.32 |
| Fe3O4 | green_synthesis | well_diffusion | hexagonal | 500.00 | 6.00 | 18.05 | 24.28 | 6.23 |
| ZnO | chemical_synthesis | well_diffusion | nanoplates | 200.00 | 8.00 | 13.69 | 19.74 | 6.05 |
| Ag | chemical_synthesis | disk_diffusion | Spherical | 500.00 | 9.00 | 18.23 | 24.18 | 5.95 |

Table S15. Reproduction of experimental results conducted by Pannerselvam et al. [1] and Sharma et al. [2] using ML reinforced GA. The rows are sorted by the percentage difference in the ascending order.

| Bacteria | NP concentration (mg/ml) | Actual ZOI (mm) | Predicted ZOI (mm) | Absolute difference | Percentage difference | Ref. |
|---|---|---|---|---|---|---|
| *P. aeruginosa* | 0.01 | 20.30 | 20.20 | 0.10 | 0.49 | [1] |
| *B. subtilis* | 10.00 | 17.70 | 17.20 | 0.50 | 2.82 | [2] |
| *E. coli* | 0.01 | 17.00 | 17.80 | 0.80 | 4.71 | [1] |
| *S. aureus* | 0.01 | 15.60 | 16.60 | 1.00 | 6.41 | [1] |
| *B. subtilis* | 0.01 | 14.60 | 15.60 | 1.00 | 6.85 | [1] |
| *B. subtilis* | 25.00 | 22.10 | 19.40 | 2.70 | 12.22 | [2] |
| *S. aureus* | 10.00 | 6.70 | 8.20 | 1.50 | 22.39 | [2] |
| *E. coli* | 10.00 | 7.70 | 9.60 | 1.90 | 24.68 | [2] |
| *E. coli* | 25.00 | 9.60 | 12.40 | 2.80 | 29.17 | [2] |
| *S. aureus* | 25.00 | 8.60 | 11.30 | 2.70 | 31.40 | [2] |
| *K. pneumoniae* | 0.01 | 14.00 | 18.60 | 4.60 | 32.86 | [1] |

Table S16. Optimized parameters of CatBoost and XGB regression models in MC dataset.

| CatBoost Regressor hyperparameter | | XGB Regressor hyperparameter | |
|---|---|---|---|
| learning_rate | 0.23630161689686982 | learning_rate | 0.006699065156114999 |
| n_estimators | 774 | n_estimators | 850 |
| depth | 5 | max_depth | 15 |
| min_child_samples | 6 | min_child_weight | 5 |
| border_count | 178 | gamma | 0.006699065156114999 |
| subsample | 0.7205585920561297 | subsample | 0.9008345277446831 |
| colsample_bylevel | 0.9246560514791413 | colsample_bytree | 0.9112361160548492 |
| l2_leaf_reg | 5.567508702983153 | reg_lambda | 9.314651052782912 |
| random_strength | 0.4857958992981014 | reg_alpha | 0.49474461625741795 |

Table S17. Optimized parameters of CatBoost and XGB regression models in ZOI dataset.

| CatBoost Regressor hyperparameter | | XGB Regressor hyperparameter | |
|---|---|---|---|
| learning_rate | 0.24258580818406097 | learning_rate | 0.29778964486473575 |
| n_estimators | 516 | n_estimators | 950 |
| depth | 7 | max_depth | 6 |
| min_child_samples | 5 | min_child_weight | 9 |
| border_count | 254 | gamma | 0.4730231011638504 |
| subsample | 0.6164651038373805 | subsample | 0.9963430887980652 |
| colsample_bylevel | 0.6486977489148302 | colsample_bytree | 0.7623101361943776 |
| l2_leaf_reg | 5.70030535531921 | reg_lambda | 2.613320937293194 |
| random_strength | 0.86690749881689924 | reg_alpha | 1.0604193051539161 |

Table S168. Comparison of CatBoost and XGB model performance for MC dataset.

| Evaluation metrics | CatBoost regressor | XGB regressor |
|---|---|---|
| R-squared score | 0.82 | 0.80 |
| Mean absolute error (MAE) | 0.31 | 0.34 |
| Mean square error (MSE) | 0.21 | 0.23 |
| Root mean square error (RMSE) | 0.46 | 0.48 |

Table S179. Comparison of CatBoost and XGB model performance for ZOI dataset.

| Evaluation metrics | CatBoost regressor | XGB regressor |
|---|---|---|
| R-squared score | 0.84 | 0.83 |
| Mean absolute error (MAE) | 1.72 | 1.78 |
| Mean square error (MSE) | 5.82 | 6.13 |
| Root mean square error (RMSE) | 2.41 | 2.48 |

**Section D. References**

1. Pannerselvam, B., Alagumuthu, T. S., Cinnaiyan, S. K., Al-Dhabi, N. A., Ponmurugan, K., Saravanan, M., Kanth, S. v., & Thangavelu, K. P. (2021). In vitro Cytotoxicity and Antibacterial Activity of Optimized Silver Nanoparticles Against Wound Infectious Bacteria and Their Morphological Studies. *Journal of Cluster Science*, *32*(1), 63–76.
2. Sharma, S., & Kumar, K. (2021). Aloe-vera leaf extract as a green agent for the synthesis of CuO nanoparticles inactivating bacterial pathogens and dye. *Journal of Dispersion Science and Technology*, 1–13.