Supporting Information

Predicting Uranium in Punjab's Groundwater Using Common Anions: A Machine Learning Approach

Sankar Sudhir[ψ1], Vamanie Perumal[ψ2], Tanmayaa Nayak[1], Thalappil Pradeep*[1,3]

[1]DST Unit of Nanoscience (DST UNS) and Thematic Unit of Excellence (TUE), Department of Chemistry, Indian Institute of Technology Madras, Chennai 600 036, India

[2]Department of Engineering Design, Indian Institute of Technology Madras, Chennai 600 036, India

*Email: pradeep@iitm.ac.in

Tel.: +91-44 2257 4208; Fax: +91-44 2257 0545/0509

[3]International Centre for Clean Water, 2nd Floor, B-Block, IIT Madras Research Park, Kanagam Road, Taramani, Chennai 600113, India

Contents

Total number of pages: 28

Total number of figures: 20

Total number of tables: 4

ψ These authors contributed equally to this work.

Table of Contents

Section 1: Modeling Methods

1. Decision Tree

Decision trees are non-parametric supervised learning algorithms used for classification and regression. They partition the feature space into axis-aligned rectangular regions by iteratively splitting input variables based on information gain or impurity measures (such as Gini index or entropy). Their transparent hierarchical structure enables straightforward interpretation of feature

importance and prediction logic. However, single decision trees are prone to overfitting and may perform inconsistently if the training data is noisy or unbalanced.

2. Ensemble Learning

Ensemble learning aggregates the predictions of multiple base models to improve generalization and reduce individual model variance. Approaches such as bagging, boosting, and stacking leverage the diversity or complementary strengths of individual learners to deliver more robust and accurate predictions than could be achieved by any single constituent model.

3. Random Forest

Random Forest is an ensemble method that constructs a multitude of decision trees using bootstrapped samples and random feature subsets at each split, promoting model diversity. The ensemble's output is computed by averaging the predictions (regression) or taking a majority vote (classification). This approach significantly reduces overfitting, handles high-dimensional and collinear features effectively, and provides meaningful estimations of feature importance.

4. Gradient Boosting

Gradient Boosting builds additive models in a forward stage-wise manner by sequentially fitting weak learners (typically shallow trees) to the residual errors from prior models. Each iteration aims to minimize a specified loss function using gradient descent, resulting in a strong composite model capable of capturing complex, non-linear relationships. While highly accurate, gradient boosting methods must be carefully regularized to avoid overfitting.

5. XGBoost

XGBoost is an optimized implementation of gradient boosting that incorporates both L1 and L2 regularization, efficient tree construction, parallelization, and robust handling of missing values. It is particularly noted for its computational efficiency, scalability to large datasets, and superior predictive performance, making it a popular choice for a wide range of structured data applications.

6. K-means Clustering

K-means is an unsupervised learning algorithm that partitions n samples into k clusters by iteratively assigning points to the nearest centroid and then recalculating centroids to minimize within-cluster variance. While it is a fast and simple algorithm, performance is sensitive to feature scaling and the initial placement of centroids, and it assumes clusters are convex and isotropic in shape.

7. Gaussian Mixture Model (GMM) Clustering

The Gaussian Mixture Model is a probabilistic clustering technique that represents the data as a mixture of multiple Gaussian distributions. Unlike K-means, GMM allows soft assignment of points, meaning each point is associated with a probability of belonging to each cluster. This approach can model clusters of varied shapes and densities and is more flexible for overlapping or non-convex clusters.

8. Residual-based Clustering

Residual-based clustering involves analyzing the residuals (errors) from a baseline predictive model. By clustering these residuals, one can identify subgroups within the data where the model underperforms. Subsequent local models are then trained within such regions to correct systematic errors, thereby improving overall predictive reliability in heterogeneous datasets.

9. Weighted Averaging Ensemble

In weighted averaging, predictions from several candidate models are combined, with weights assigned inversely proportional to validation error or directly proportional to model accuracy. This strategy enhances robustness by placing greater emphasis on well-performing predictors, yielding an aggregate result less sensitive to the weaknesses of any single model.

10. Stacking

Stacking is a hierarchical ensemble method where first-level (base) models are trained independently, and their predictions are then used as input features for a second-level "meta-learner." The meta-model learns optimal combinations of base predictions, typically using out-of-fold estimates to avoid data leakage, and thus can exploit complex inter-model relationships for improved performance.

11. Voting for Regression

In regression tasks, voting (typically simple or weighted averaging) combines predictions from multiple base models to produce a single output. Averaging smooths out random fluctuations from individual estimators, and is particularly effective when constituent models are diverse but moderately accurate.

Section S2: Model Evaluation Metrics:

1. Coefficient of Determination (R^2):

The coefficient of determination, denoted as R^2, measures the degree to which the predicted values match the actual observations. It quantifies the proportion of variance in the dependent variable (observed uranium concentrations) that is explained by the independent variables (model

predictions). An R^2 value closer to 1 indicates a better fit, meaning the model can explain most of the variability in the data. For example, an R^2 value of 0.9 implies that 90% of the variance in observed values is captured by the model, leaving only 10% unexplained.

2. Root Mean Square Error (RMSE):

RMSE represents the standard deviation of residuals (the differences between observed and predicted values). It provides a measure of how spread out these errors are and is calculated as the square root of the average squared differences between observed and predicted values. A lower RMSE indicates better model performance, as it reflects smaller deviations from actual observations. RMSE is particularly sensitive to large errors due to its squared term, making it an important metric for identifying significant prediction inaccuracies.

3. Mean Absolute Error (MAE):

MAE calculates the average magnitude of absolute differences between observed and predicted values. Unlike RMSE, MAE treats all errors equally without squaring them, making it less sensitive to large outliers. It provides a straightforward measure of prediction accuracy in the same units as the data. A smaller MAE indicates a more accurate model, with fewer deviations from observed values on average.

4. Correlation Coefficient (r):

The correlation coefficient measures the strength and direction of the linear relationship between observed and predicted values. It ranges from -1 to +1, where a value closer to +1 indicates a strong positive correlation (as one variable increases, so does the other), while a value closer to -1 indicates a strong negative correlation. A value near 0 suggests no linear relationship. The

correlation coefficient complements R^2 by providing insight into how closely related the two variables are in terms of their linear association.

Table S1. Financial year–wise (April 1 to March 31) number of drinking-water schemes tested (tubewell/canal/handpump) in the DWSS monitoring program (2018–19 to 2023–24).

Financial Year	Total No. of schemes tested			
	Tubewell	Canal	Handpump	Total
2018-19	2609	69	339	3017
2019-20	1860	81	222	2163
2020-21	1485	17	112	1614
2021-22	3143	445	126	3714
2022-23	5641	718	348	6707
2023-24	5806	777	290	6873

Table S2. Descriptive statistics of the observed water quality parameters in Punjab with 8735 data points.

Feature	IQR Outliers	Z-score Outliers	Min	Max	Mean	Std
Lead	191	92	0	0.032	0.000095	0.0008
Chromium	243	136	0	0.045	0.000132	0.0010
Mercury	196	35	0	0.009	0.000020	0.0002
Arsenic	1372	190	0	0.094	0.001925	0.0073
Cadmium	138	3	0	2.577	0.000565	0.0321

Nickel	184	47	0	0.06	0.000097	0.0013
Iron	933	85	0	4.068	0.034102	0.1287
Fluoride	563	160	0	9.83	0.553128	0.5607
Chloride	1068	135	0	890.2	27.930558	55.4283
Nitrate	1035	106	0	240.56	6.575983	12.9697
Sulphate	957	293	0.09900	399.86	50.822982	70.5093
Uranium	315	159	0.00001	335.11	17.292389	21.2731

Table S3. Comparison of validation metrics for clustering.

Method	**Silhouette**	**Davies-Bouldin**	**Calinski-Harabasz**	**BIC**
K-means	0.4557	1.9505	693.86	–
GMM	0.5496	2.8541	350.50	-57628.43

The clustering analysis was conducted using both K-means and GMM with k=2, as higher k values led to unbalanced clusters with very few points in one group. Across 8,735 data points, both approaches provided reasonably balanced clusters, especially for K-means, which avoided highly skewed splits. Both clustering methods were evaluated on established validation metrics: K-means showed moderate Silhouette and higher Calinski-Harabasz scores and a suboptimal Davies-Bouldin Index, while GMM displayed a higher Silhouette but a higher Davies-Bouldin Index and a lower Calinski-Harabasz Index with strong support for k=2 from the BIC. These findings, together with balanced cluster sizes, justify the use of two clusters in this dataset.

Table S4: Mann-Whitney U test for each feature across K-means clusters.

Feature	**Mann - Whitney U Statistic**	**p** - value

Arsenic	8977249	0.000
Nitrate	7758048.5	0.000
Chloride	7788532.5	0.000
Iron	8891103.5	0.000
Sulphate	7869325	0.000
Fluoride	8036365	0.002
Mercury	8311855	0.014
Lead	8448026.5	0.018
Chromium	8454553	0.020
Cadmium	8407097.5	0.285
Nickel	8398933.5	0.527

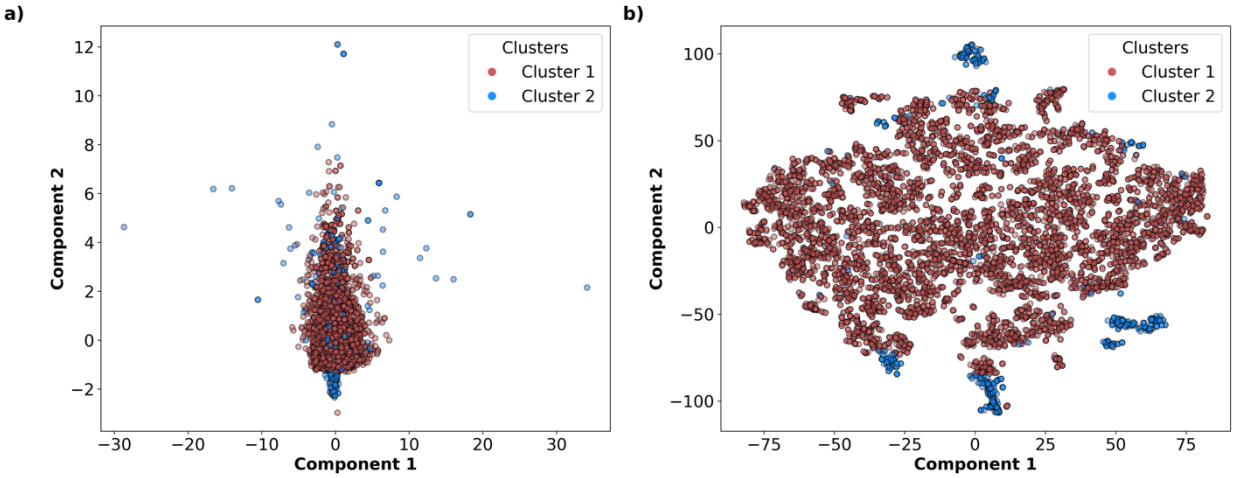


Figure S1. Visualization of clusters from GMM (a) PCA (b) t-SNE.

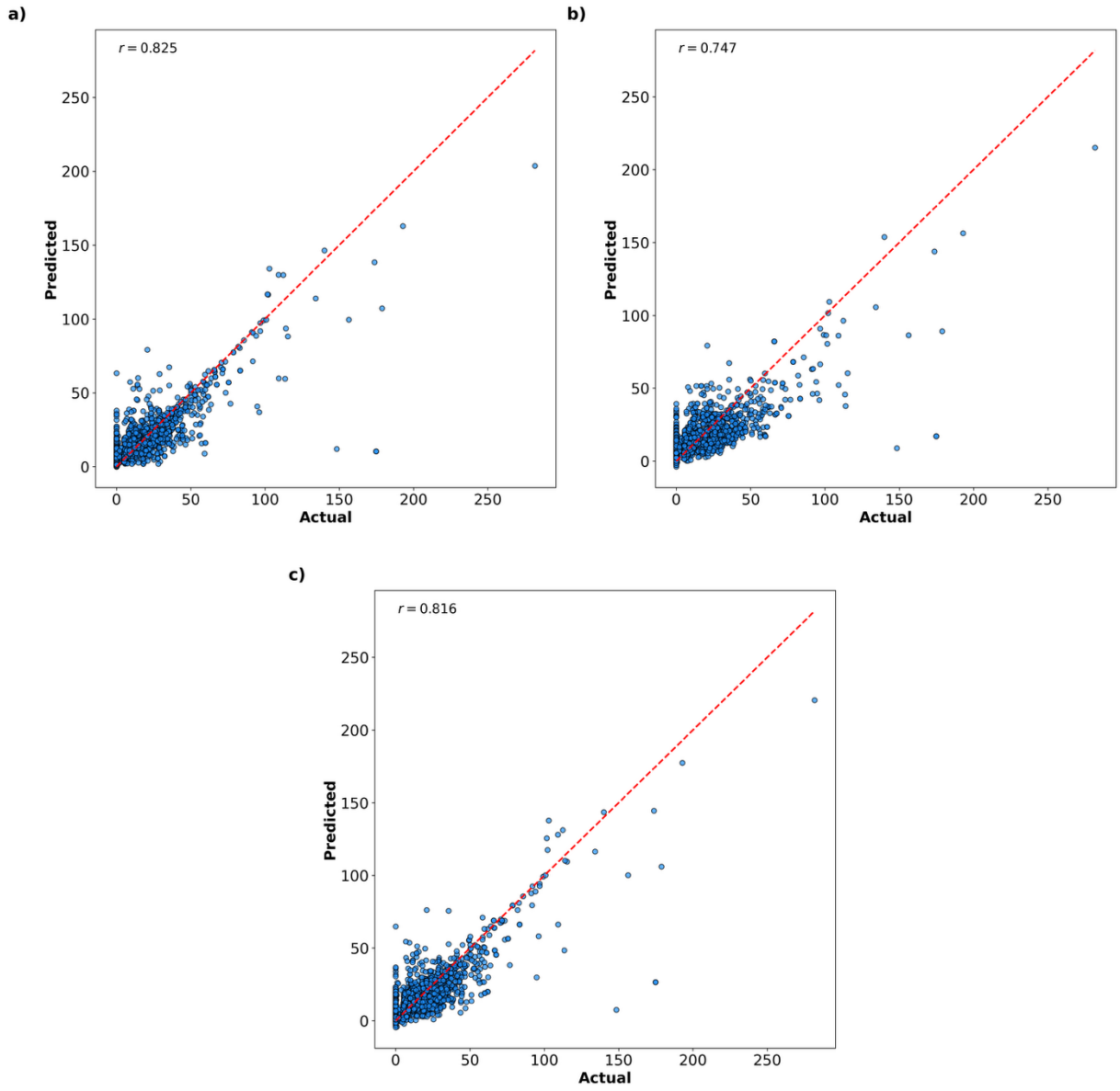


Figure S2. Actual concentration vs predicted concentration for standalone models (a) Random Forest, (b) Gradient Boosting, (c) XGBoost.

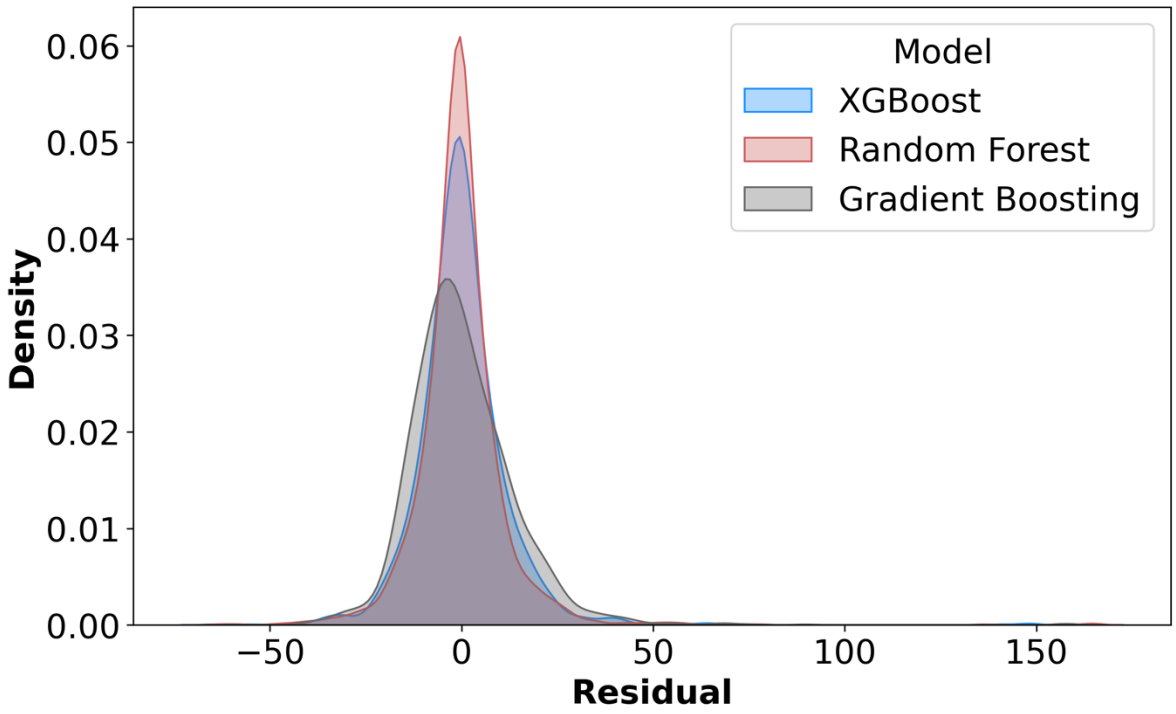


Figure S3. Residual plots for XGBoost, Gradient Boosting, and Random Forest models.

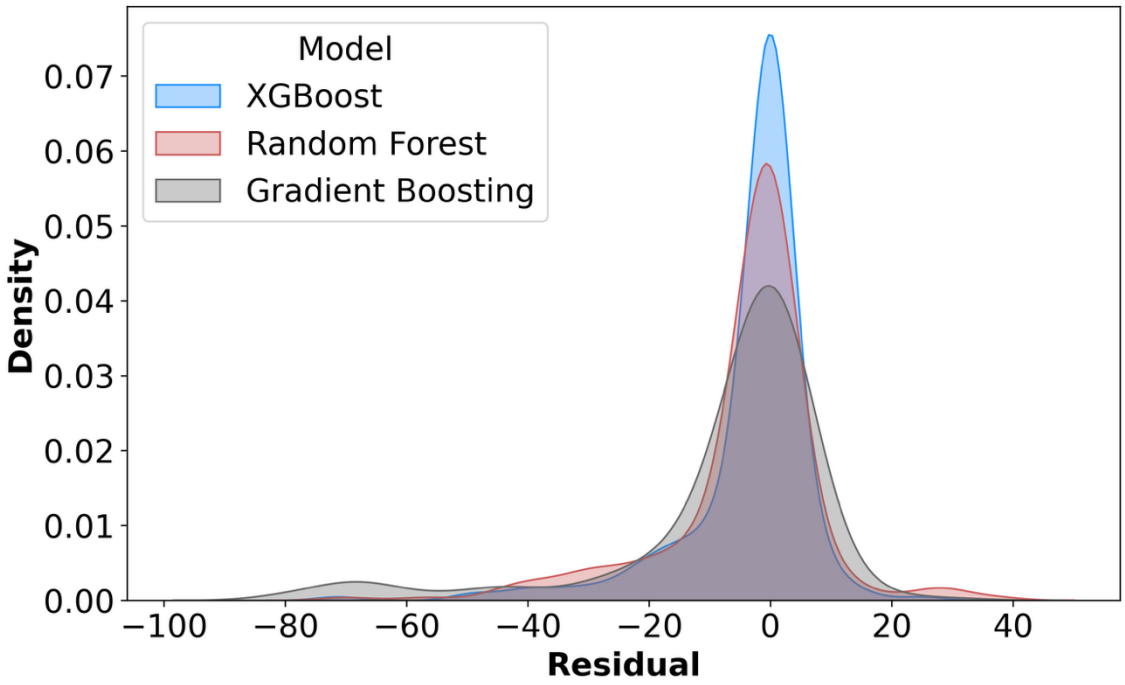


Figure S4. Residual plots for standalone models in cluster 1 (K-means).

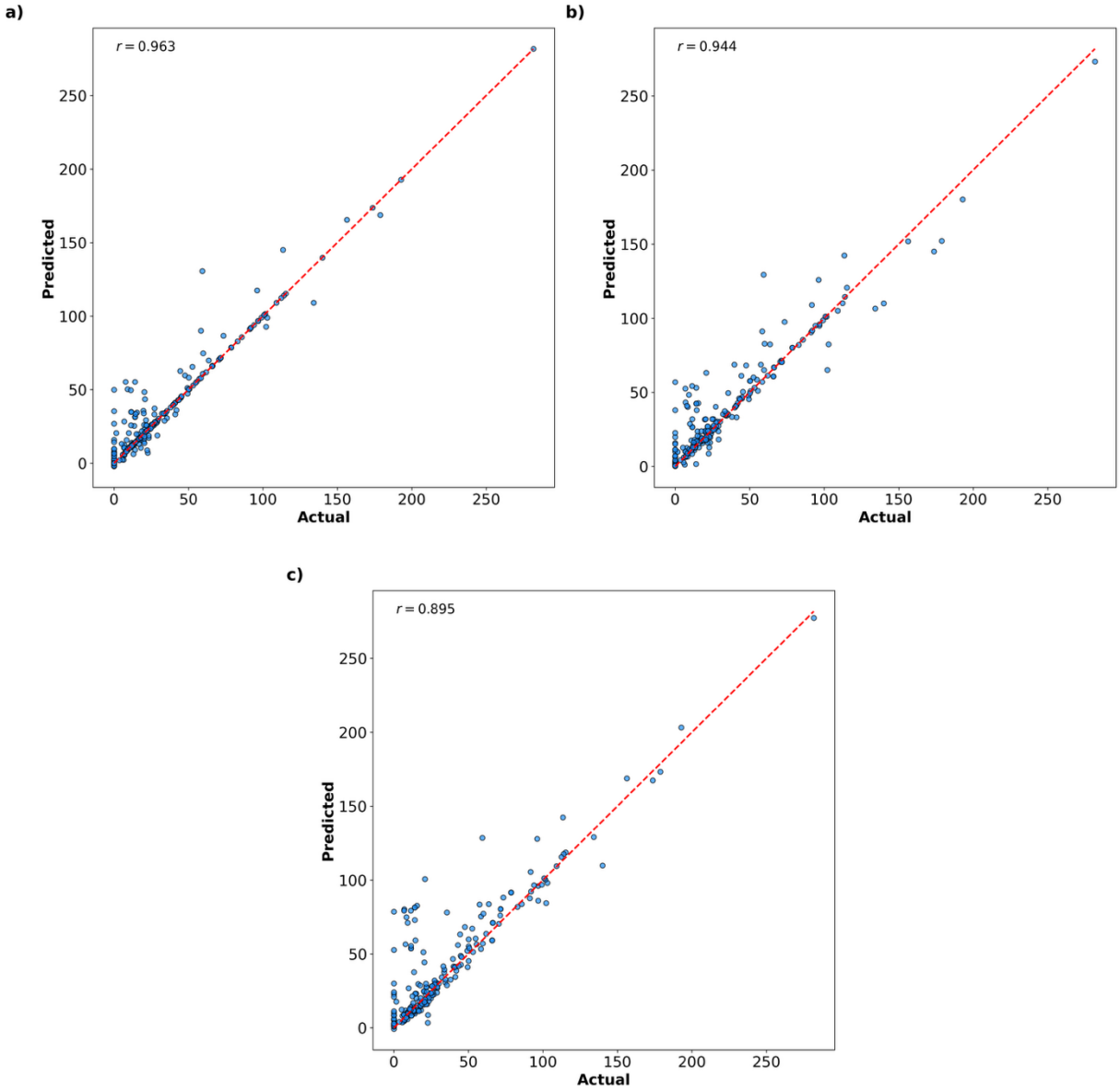


Figure S5. Cluster 1 (K-means): actual vs predicted by standalone models (a) XGBoost, (b) Random Forest (c) Gradient Boosting.

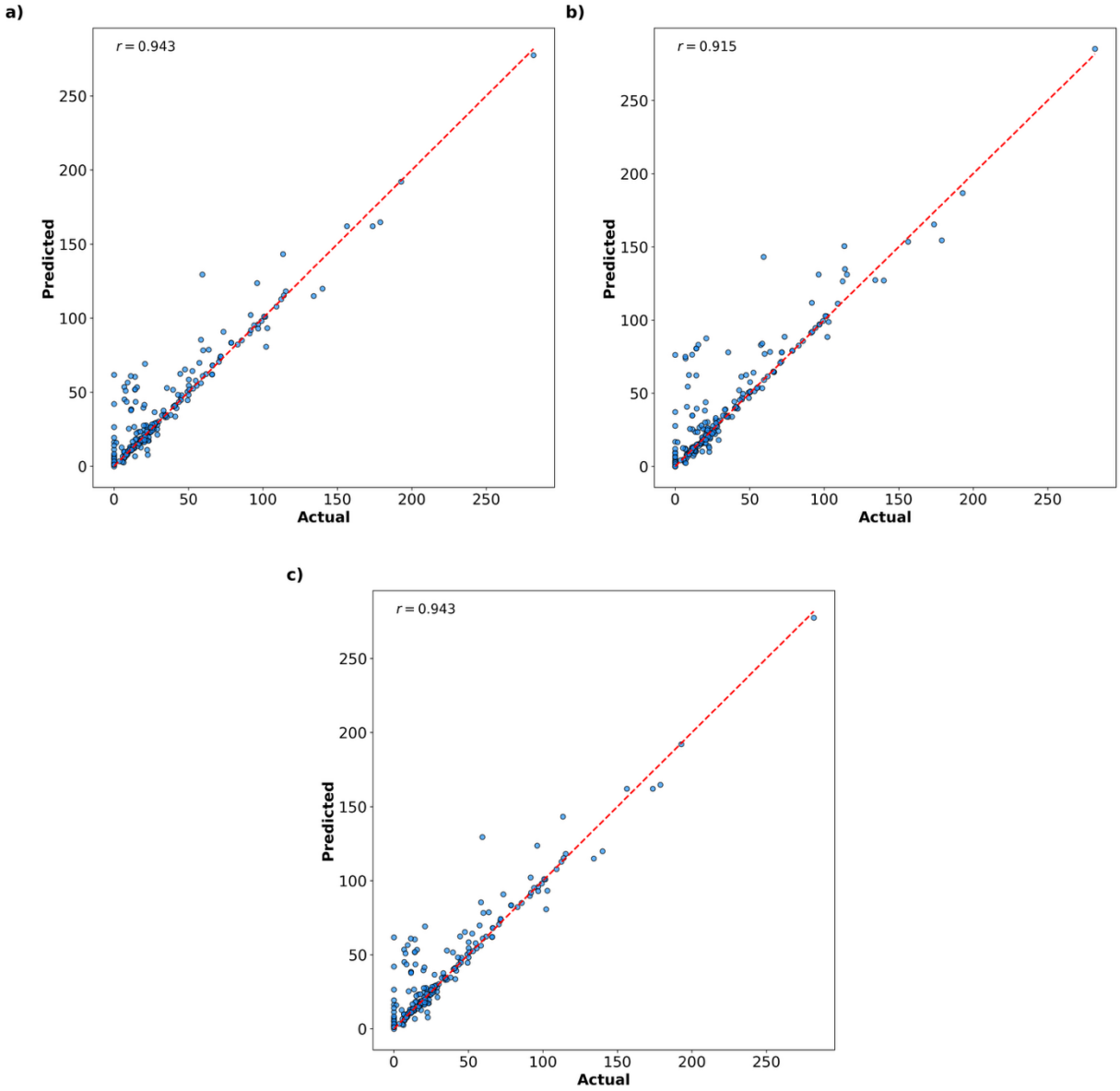


Figure S6. Cluster 1 (K-means): actual vs predicted by ensemble models (a) voting, (b) stacking, (c) weighted averaging.

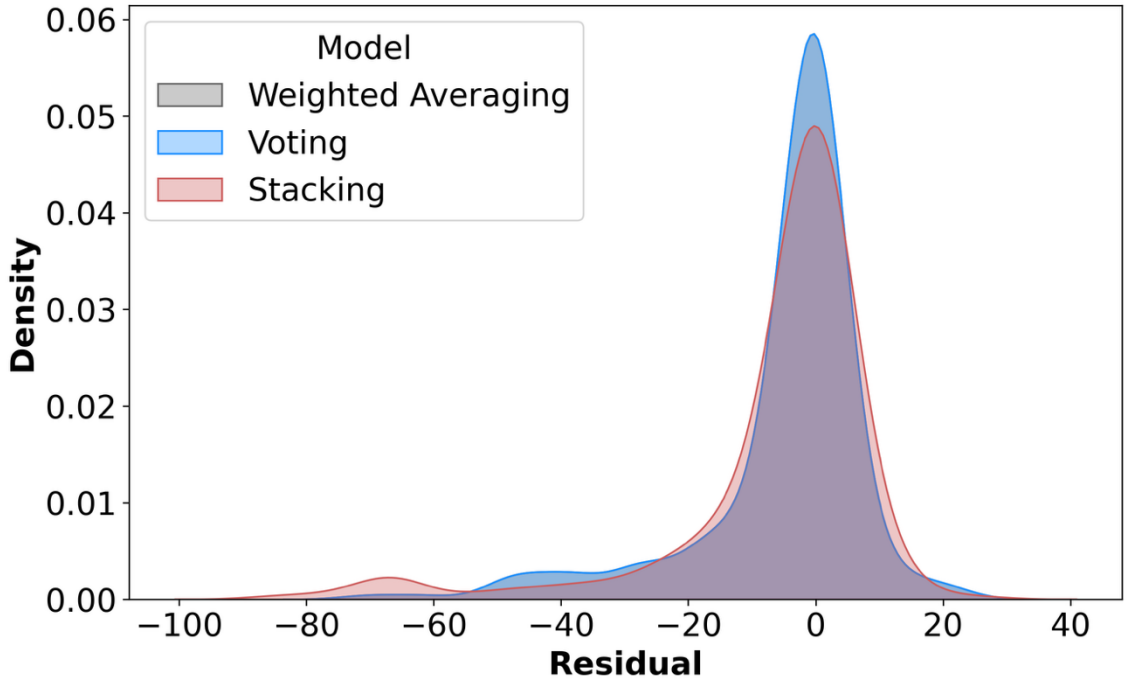


Figure S7. Residual plots for ensemble models in cluster 1 (K-means).

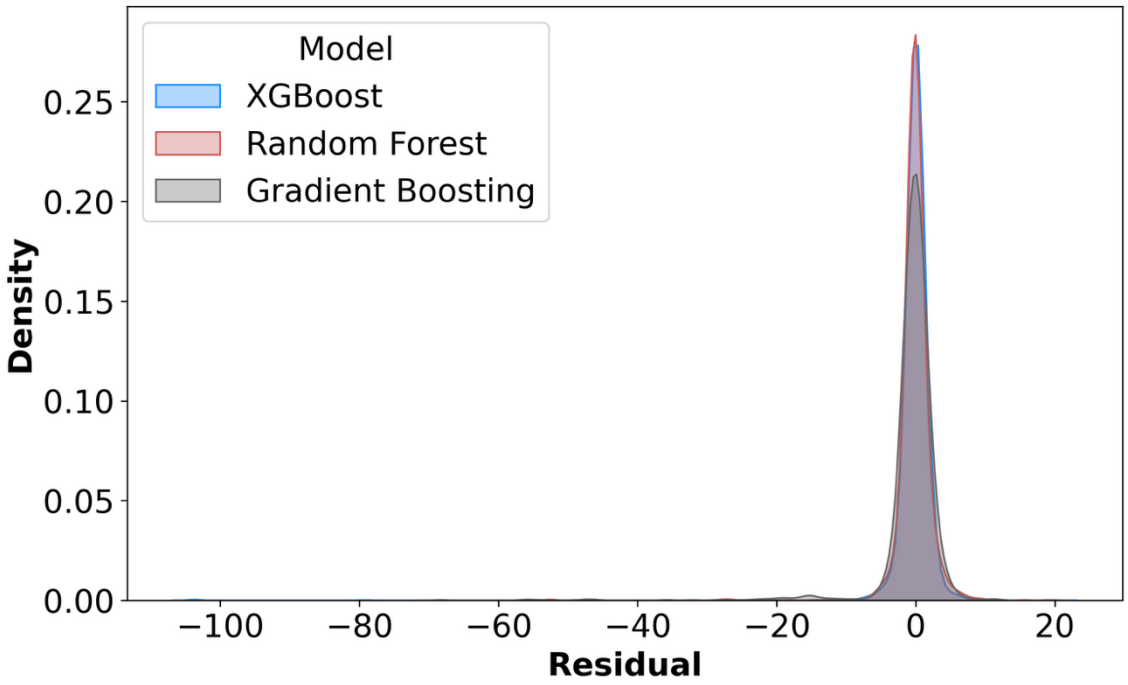


Figure S8. Residual plots for standalone models in cluster 2 (K-means).

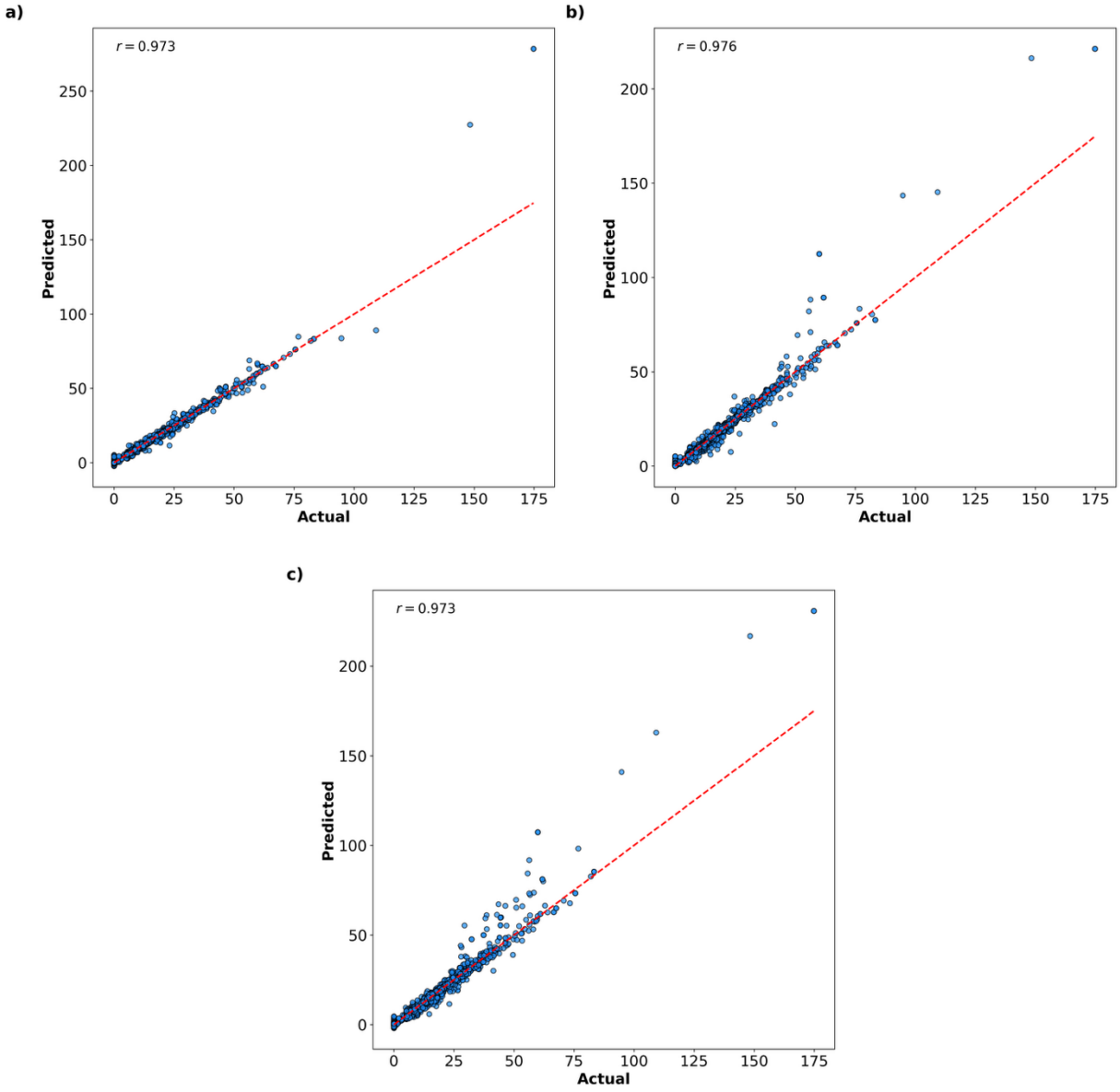


Figure S9. Cluster 2 (K-means): actual vs predicted by standalone models (a) XGBoost, (b) Random Forest, (c) Gradient Boosting.

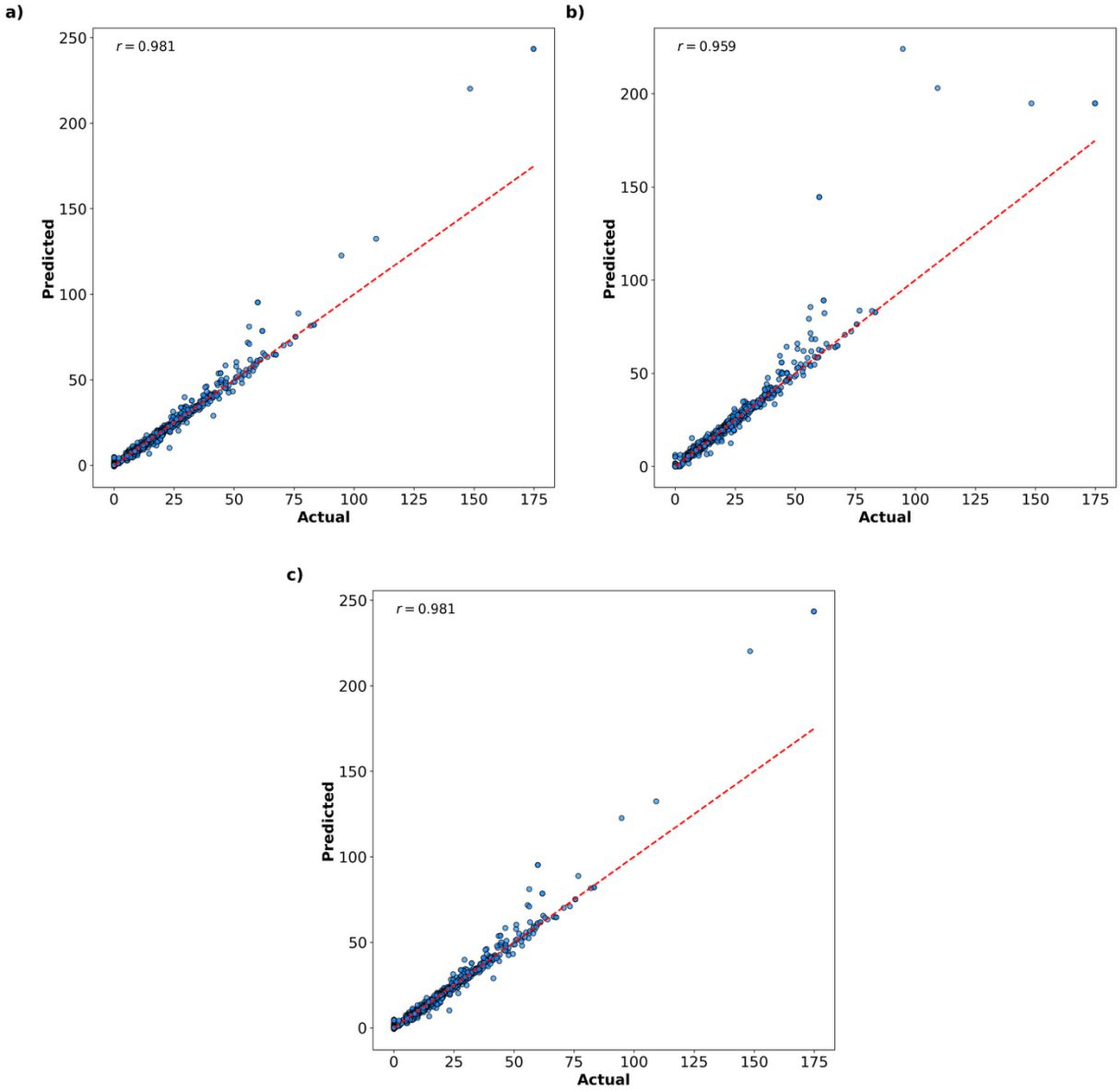


Figure S10. Cluster 2 (K-means): actual vs predicted by ensemble models (a) voting, (b) stacking, (c) weighted averaging.

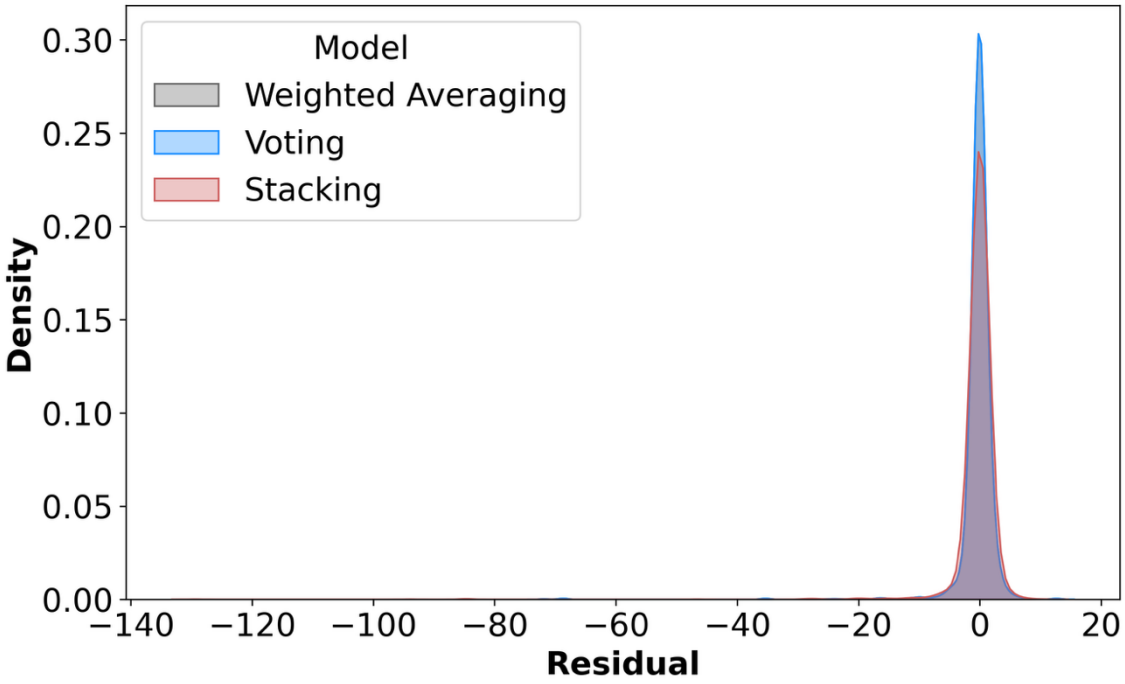


Figure S11. Residual plots for ensemble models in cluster 2 (K-means).

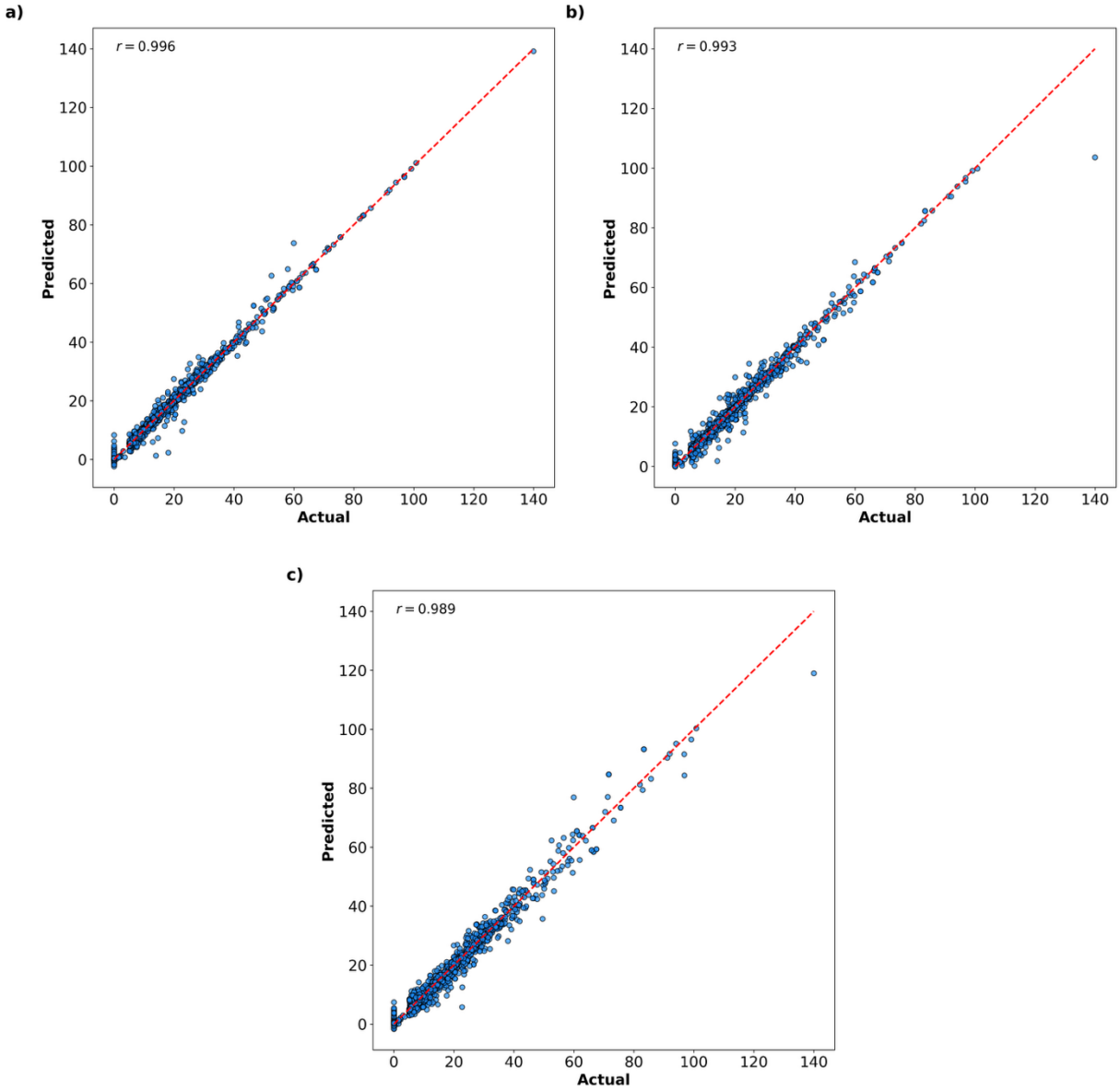


Figure S12. Cluster 1 (GMM): actual vs predicted by standalone models (a) XGBoost, (b) Random Forest, (c) Gradient Boosting.

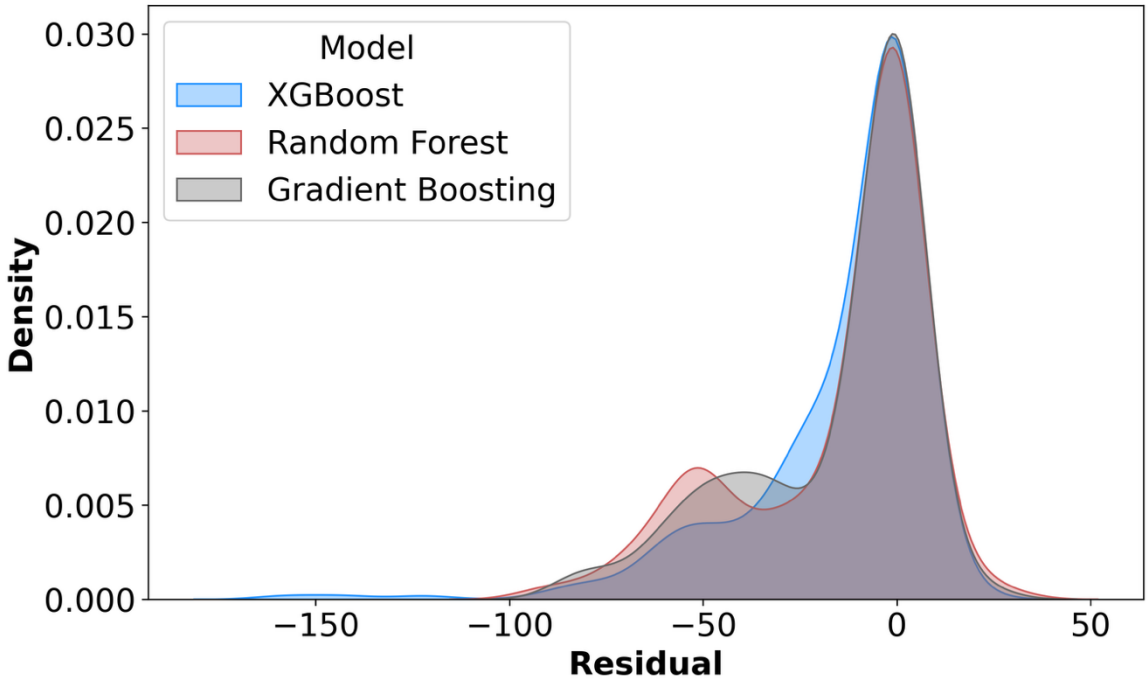


Figure S13. Residual plots for standalone models in cluster 1 (GMM).

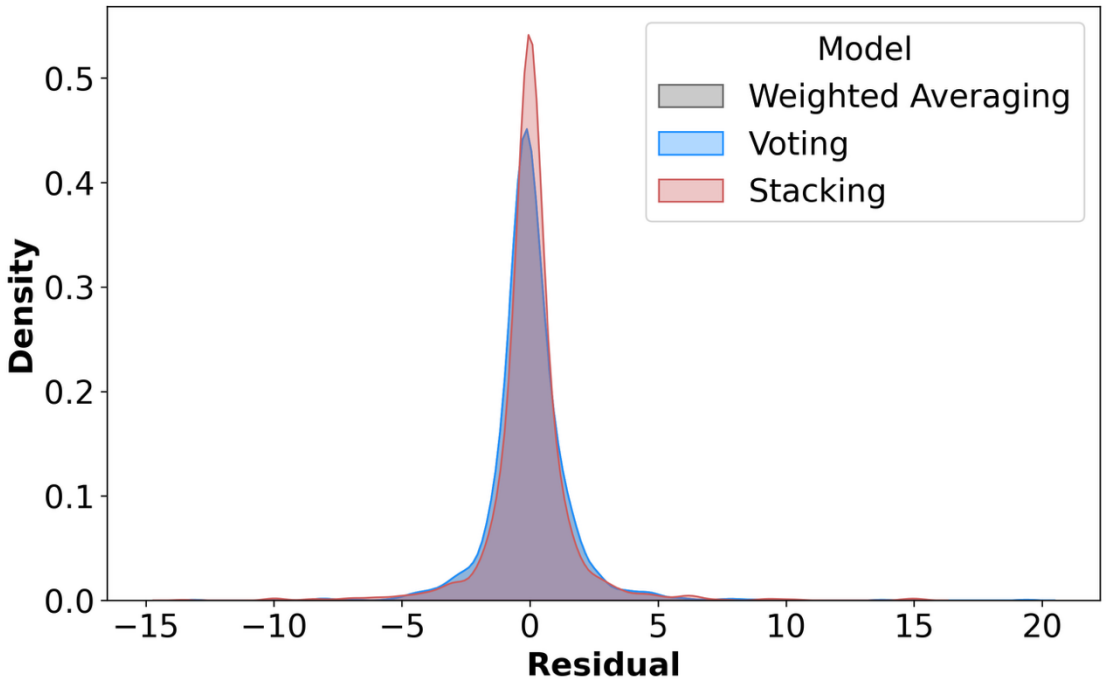


Figure S14. Residual plots for ensemble models in cluster 1 (GMM).

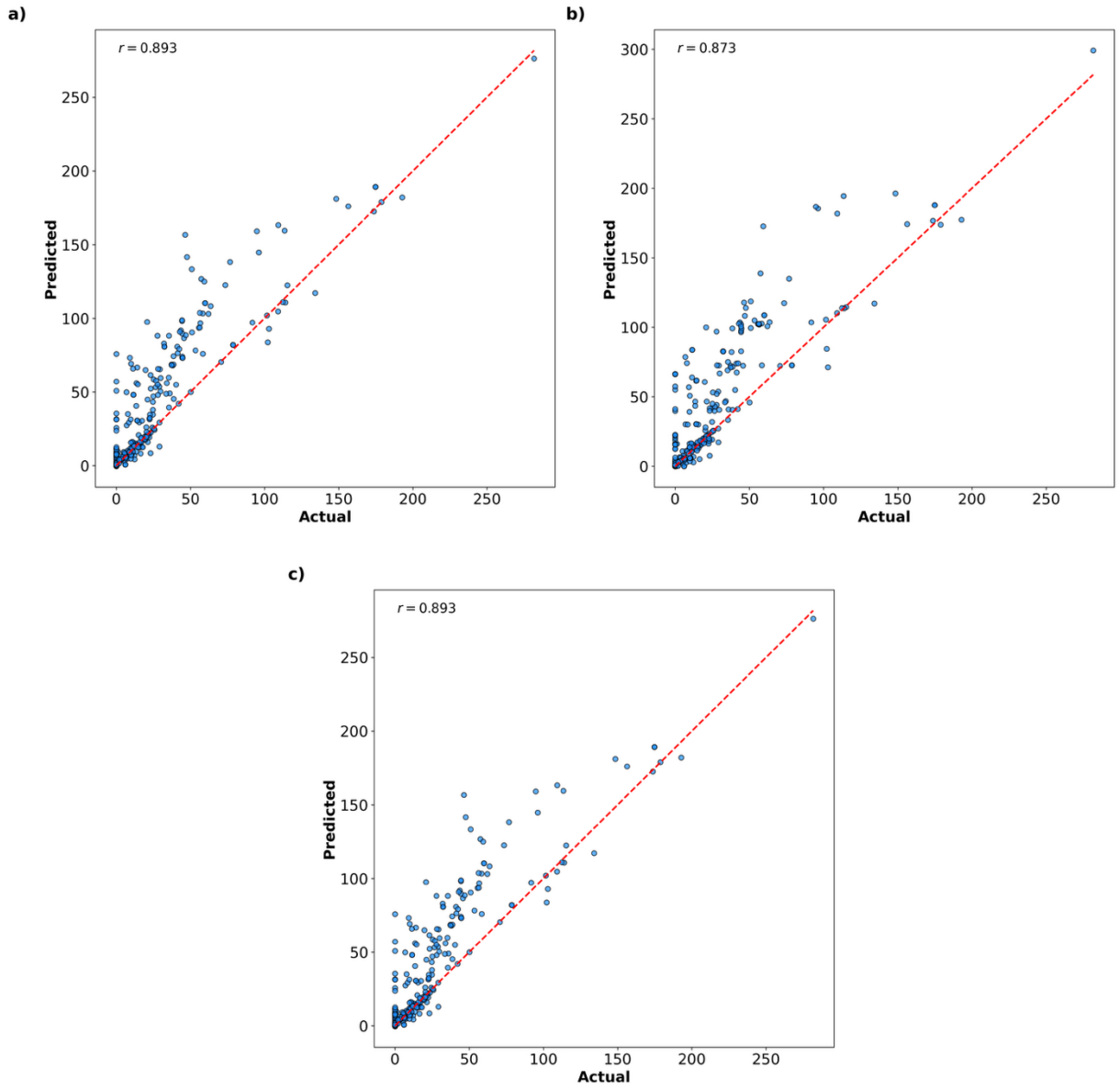


Figure S15. Cluster 1 (GMM): actual vs predicted by ensemble models (a) voting, (b) stacking, (c) weighted averaging.

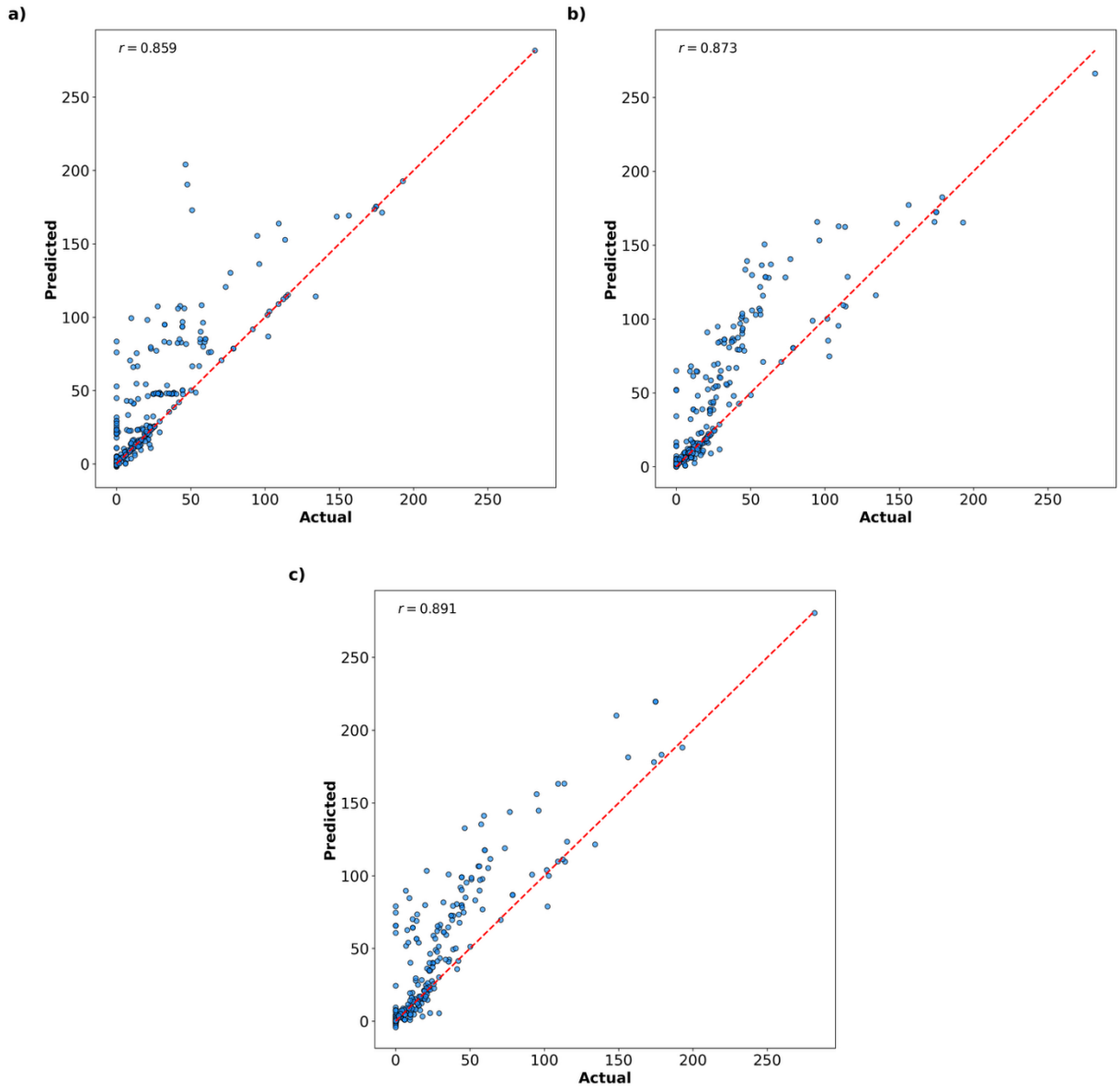


Figure S16. Cluster 2 (GMM): actual vs predicted by standalone models (a) XGBoost, (b) Random Forest, (c) Gradient Boosting.

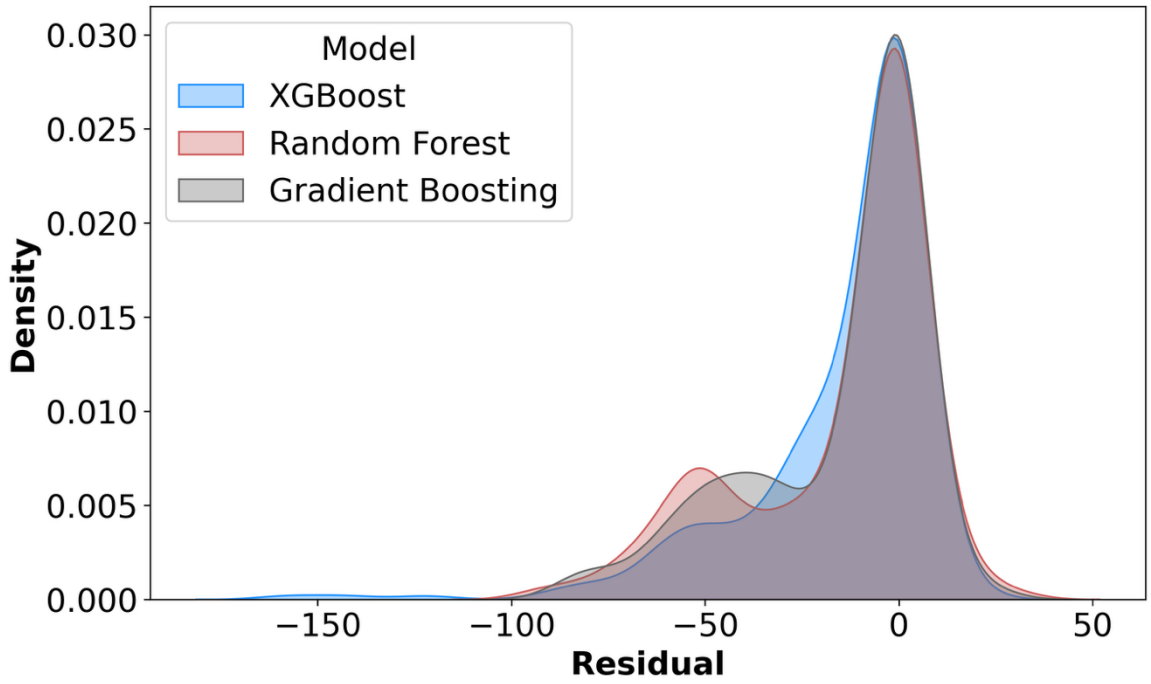


Figure S17. Residual plots for standalone models in cluster 2 (GMM).

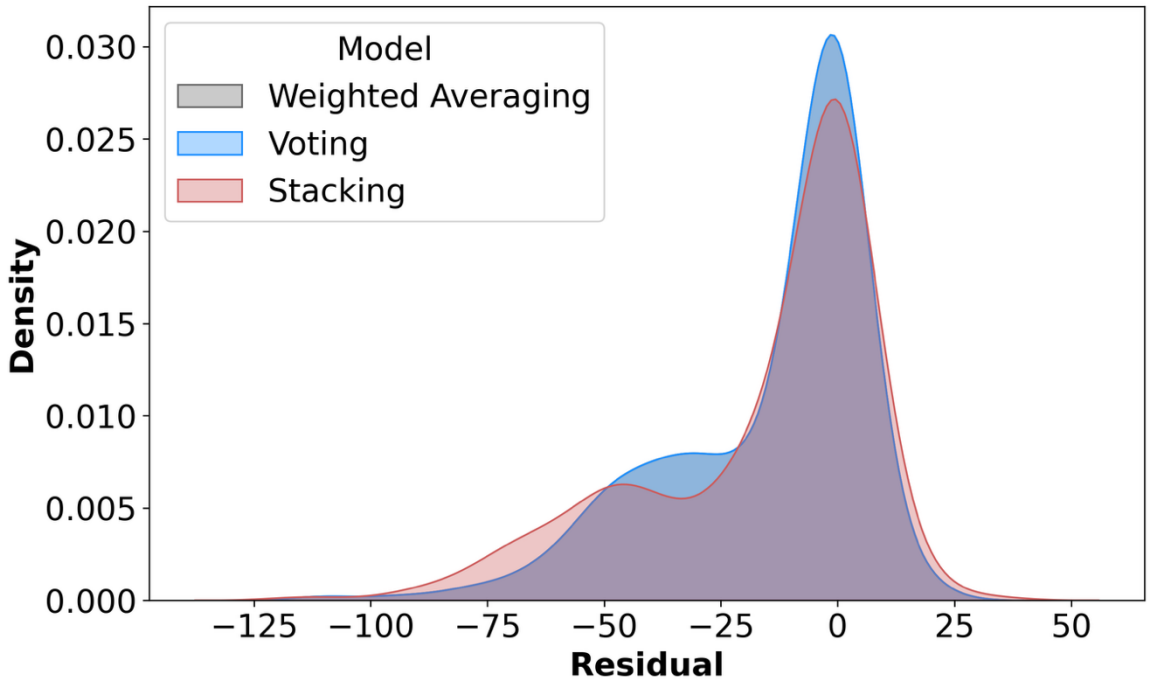


Figure S18. Residual plots for ensemble models in cluster 2 (GMM).

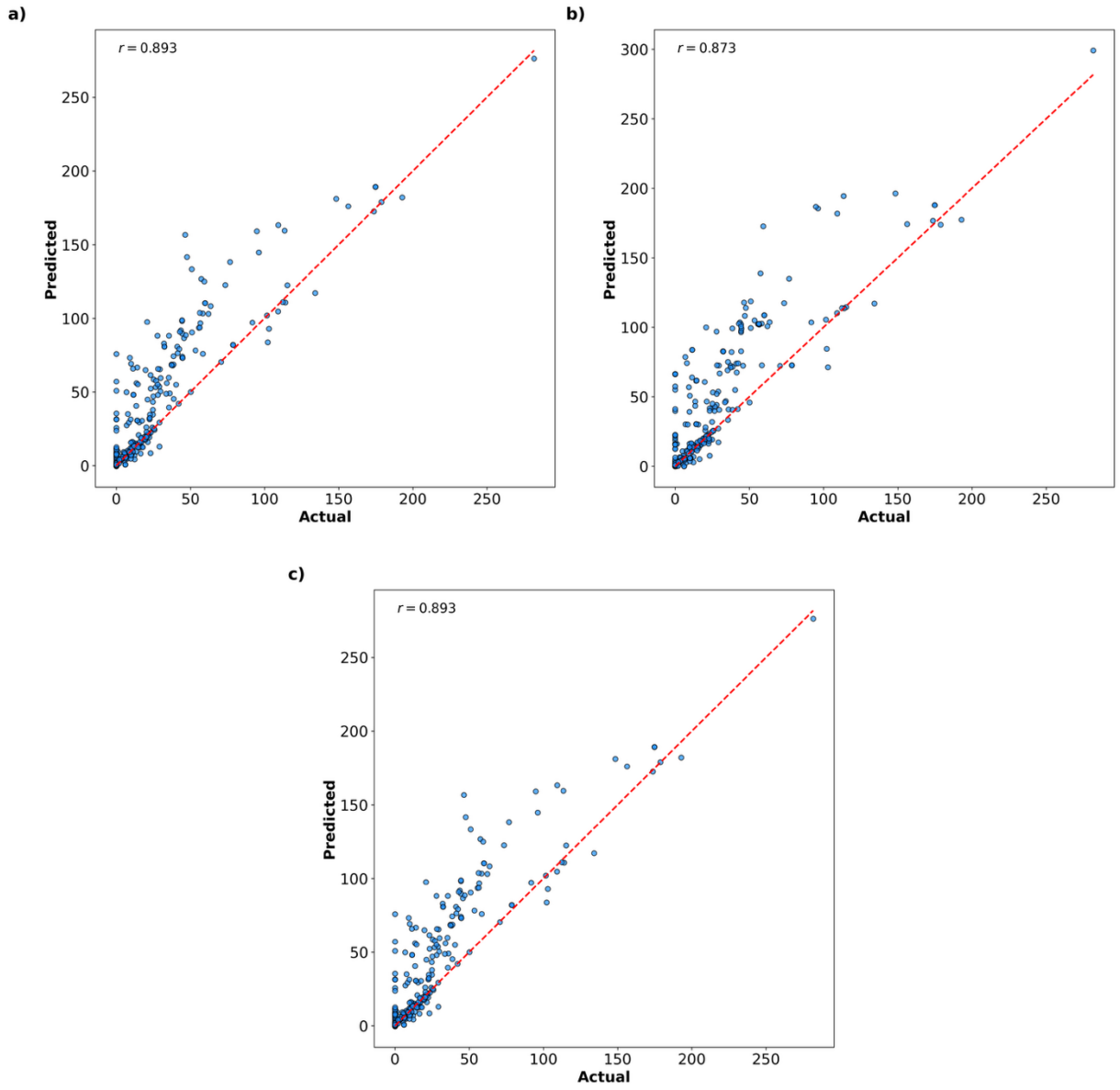
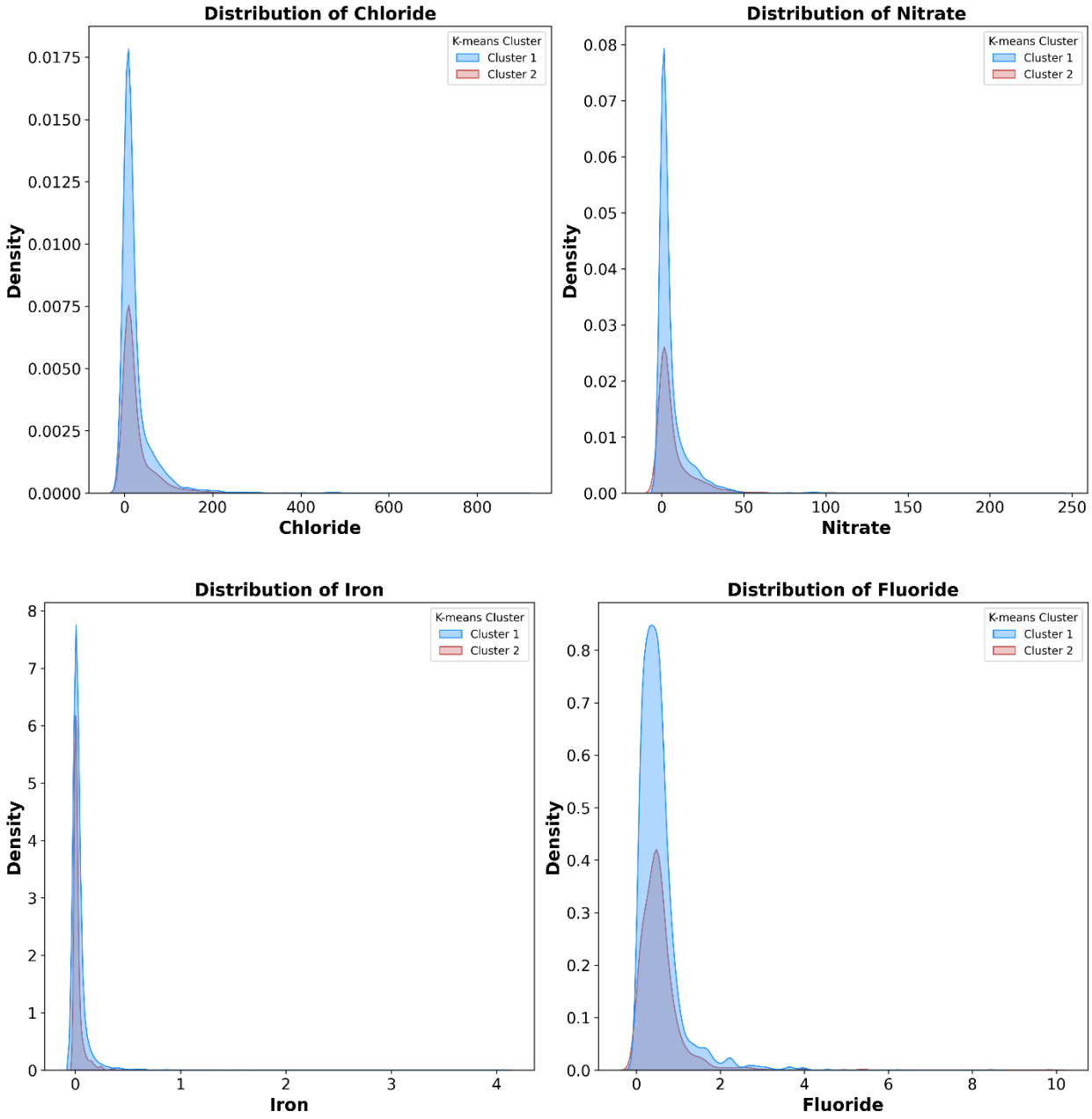
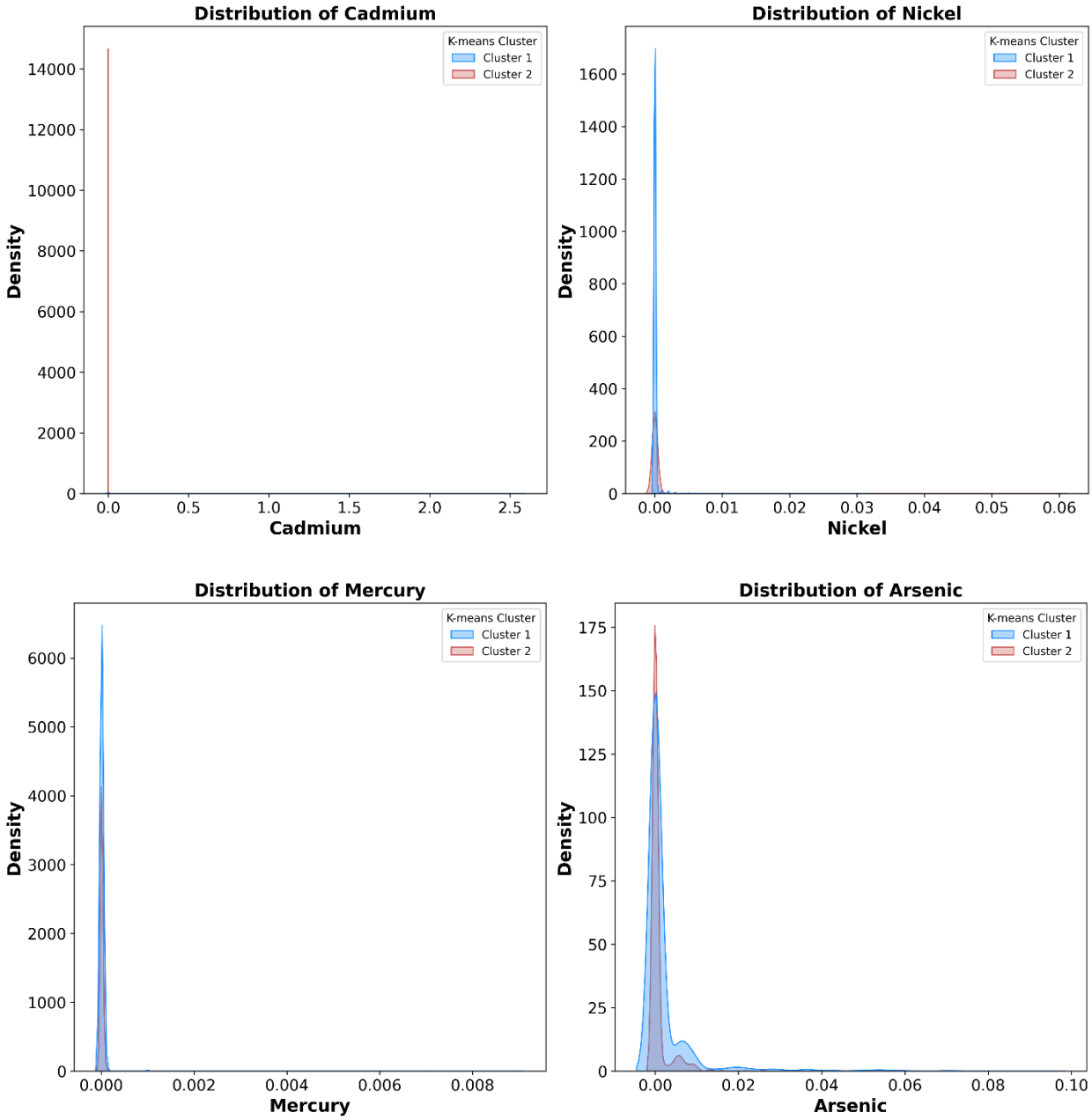


Figure S19. Cluster 2 (GMM): actual vs predicted by ensemble models (a) voting, (b) stacking, (c) weighted averaging.


Distribution of Chloride
K-means Cluster
Cluster 1
Cluster 2
Density
Chloride
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
0
200
400
600
800
Distribution of Nitrate
K-means Cluster
Cluster 1
Cluster 2
Density
Nitrate
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0
50
100
150
200
250
Distribution of Iron
K-means Cluster
Cluster 1
Cluster 2
Density
Iron
0
1
2
3
4
5
6
7
8
0
1
2
3
4
Distribution of Fluoride
K-means Cluster
Cluster 1
Cluster 2
Density
Fluoride
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
2
4
6
8
10


Distribution of Cadmium
K-means Cluster
Cluster 1
Cluster 2
Density
Cadmium
Distribution of Nickel
Nickel
Distribution of Mercury
Mercury
Distribution of Arsenic
Arsenic

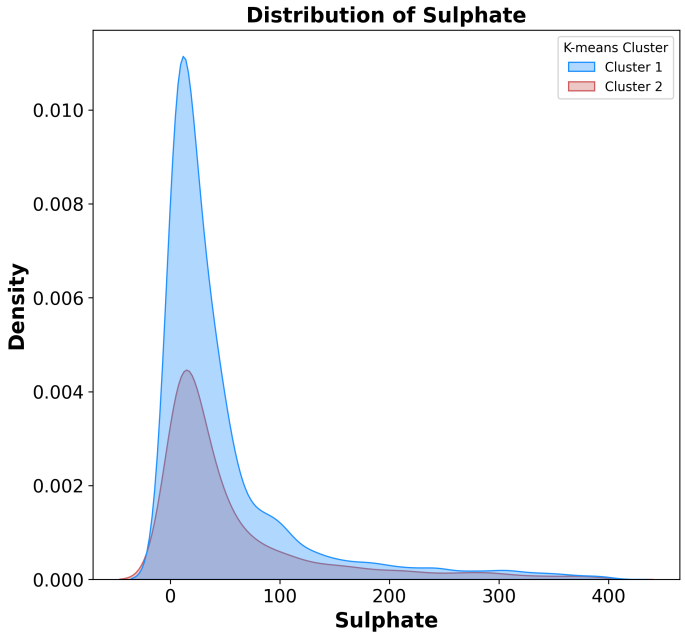


Figure S20: Statistical differences between the K-means clusters across the features.