Forecasting SARS-CoV-2 outbreak through wastewater analysis: a success in wastewater-based epidemiology

Ruben Cañas Cañas[1,2,3,4], Raimundo Seguí López-Peñalver[1], Jorge Casaña Mohedo[1,5], José Vicente Benavent Cervera[1], Julio Fernández Garrido[6], Raúl Juárez Vela[7], Ana Pellín Carcelén[1], Óscar García-Algar[3,4,8], Vicente Gea Caballero[1#], Vicente Andreu-Fernández (✉)[1,3,9 #]

1 Faculty of Health Sciences, Valencian International University (VIU), Valencia 46002, Spain

2 Global Omnium, Valencia 46005, Spain

3 Grup de Recerca Infancia i Entorn (GRIE), Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Barcelona 08036, Spain

4 Department de Cirurgia i Especialitats Mèdico-Quirúrgiques, Universidad de Barcelona, Barcelona 08036, Spain

5 Faculty of Health Sciences, Universidad Católica de Valencia San Vicente Mártir, Valencia 46001, Spain

6 Department of Nursing, University of Valencia, Valencia46001, Spain.

7 Faculty of Health Sciences, La Rioja University, Logroño 26006, Spain

8 Department of Neonatology, Instituto Clínic de Ginecología, Obstetricia y Neonatología (ICGON), Hospital Clínic-Maternitat, BCNatal, Barcelona 08028, Spain

9 Biosanitary Research Institute, Valencian International University (VIU), Valencia 46002, Spain

✉ Corresponding author
E-mail: vandreu@universidadviu.com

#These authors contributed equally as last authorship

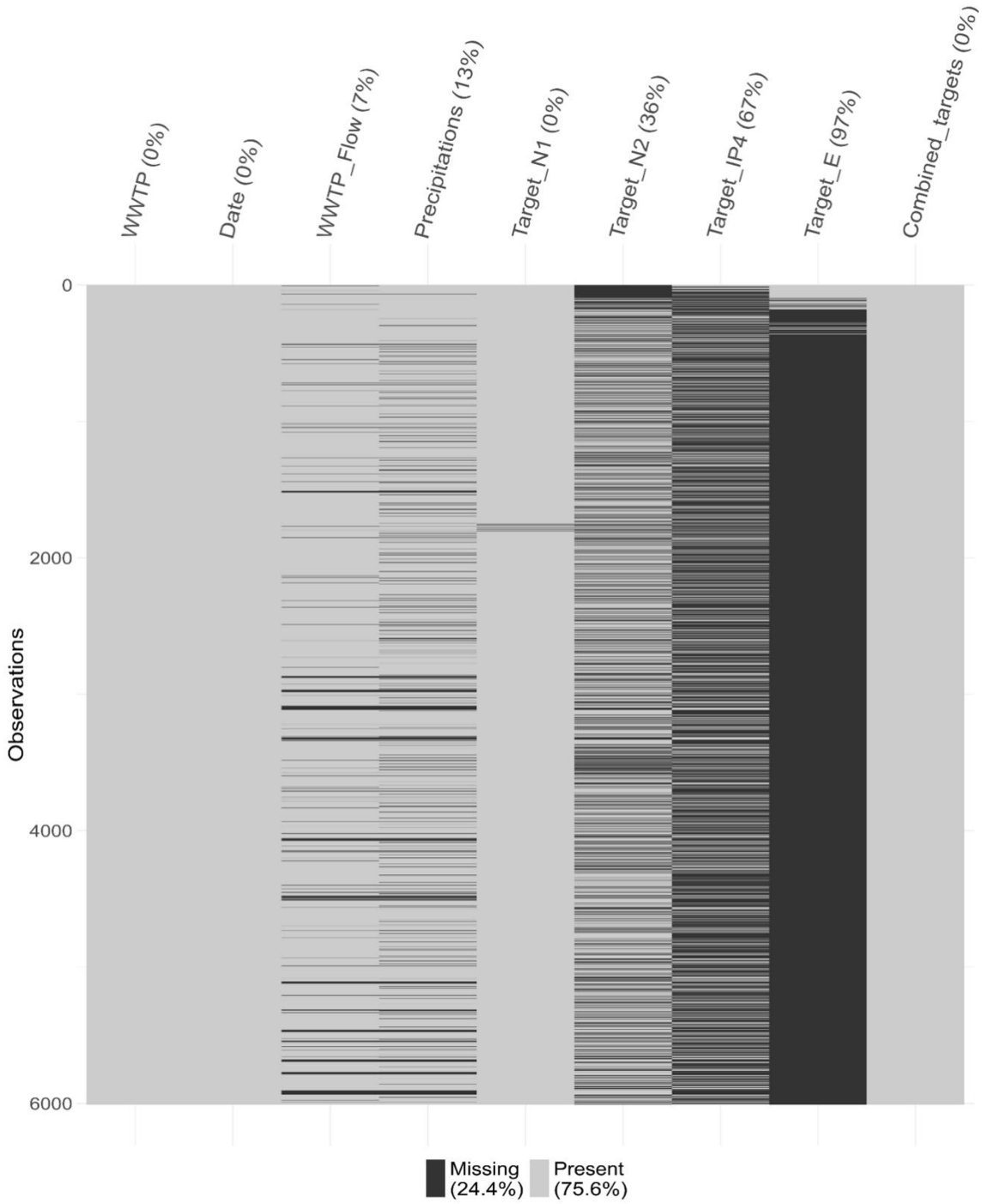

Figure S1. Representation of missing data for amplified genetic targets and covariables in wastewater treatment plant dataset. Each observation of the dataset is represented as a tile. The percentage of missing observations for each variable is represented on the figure. The combined targets variable represents the maximum value of the different gene target variables, as explained in 2.3.

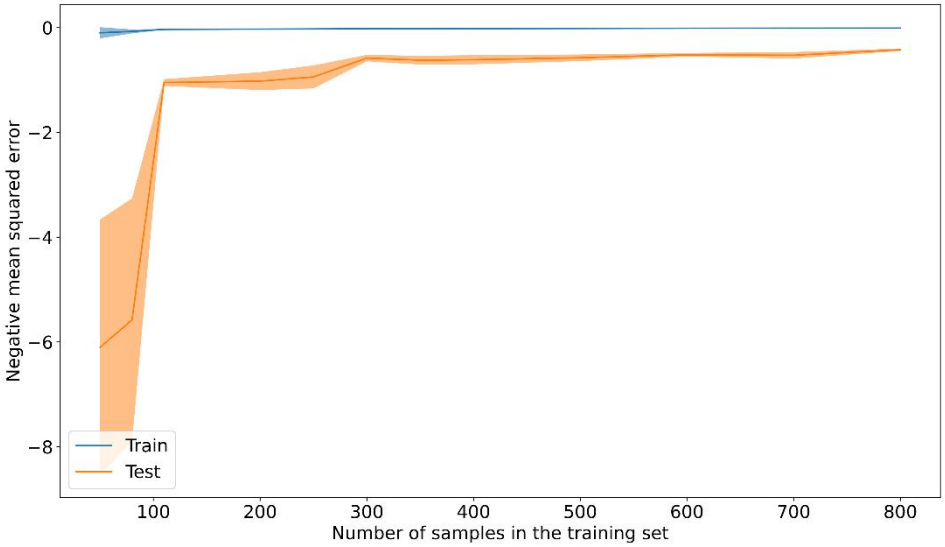

Figure S2. Learning curve of the fine-tuned Light Gradient-Bosting Model. Training score remains high regardless of the size of the training set, while test score increases with the size of the training dataset until it reaches a plateau.

Table S1. Preliminary evaluation of the models. Splits for model evaluation were 75% for the training set and 25% for the testing set. The models were assessed using both the original dataset values and the Box-Cox transformed values.

Data without transformations					Box-Cox transformed data				
Model	**Adjusted R-Squared**	**R-Squared**	**RMSE**	**Time taken**	**Model**	**Adjusted R-Squared**	**R-Squared**	**RMSE**	**Time taken**
ExtraTreesRegressor	0.84	0.84	251.53	0.54	LGBMRegressor	0.79	0.79	0.46	0.04
XGBRegressor	0.74	0.74	320.00	0.26	HistGradientBoostingRegressor	0.79	0.79	0.46	0.16
LGBMRegressor	0.67	0.67	358.63	0.08	XGBRegressor	0.79	0.79	0.46	0.23
HistGradientBoostingRegressor	0.65	0.65	370.47	0.21	ExtraTreesRegressor	0.78	0.78	0.47	0.61
KNeighborsRegressor	0.61	0.62	389.25	0.03	RandomForestRegressor	0.77	0.77	0.48	1.56
RandomForestRegressor	0.55	0.55	419.13	1.36	BaggingRegressor	0.75	0.75	0.50	0.16
MLPRegressor	0.55	0.55	421.14	1.55	MLPRegressor	0.75	0.75	0.50	1.83
TransformedTargetRegressor	0.54	0.54	424.13	0.01	GradientBoostingRegressor	0.73	0.73	0.52	0.52
LinearRegression	0.54	0.54	424.13	0.01	SVR	0.73	0.73	0.52	0.45
LassoLarsIC	0.54	0.54	424.13	0.02	NuSVR	0.73	0.73	0.52	0.40
Lars	0.54	0.54	424.13	0.01	KNeighborsRegressor	0.70	0.70	0.55	0.05
Ridge	0.54	0.54	424.16	0.01	KernelRidge	0.63	0.63	0.61	0.39
RidgeCV	0.54	0.54	424.55	0.01	TransformedTargetRegressor	0.63	0.63	0.61	0.01
BayesianRidge	0.54	0.54	424.78	0.01	LinearRegression	0.63	0.63	0.61	0.02
Lasso	0.54	0.54	424.84	0.03	Lars	0.63	0.63	0.61	0.02
LassoCV	0.54	0.54	425.55	0.11	LassoLarsIC	0.63	0.63	0.61	0.02
LarsCV	0.54	0.54	425.69	0.03	Ridge	0.63	0.63	0.61	0.01
LassoLarsCV	0.54	0.54	425.69	0.04	RidgeCV	0.63	0.63	0.61	0.01
OrthogonalMatchingPursuitCV	0.53	0.54	426.53	0.02	LassoLarsCV	0.63	0.63	0.61	0.04
LassoLars	0.53	0.54	426.59	0.01	BayesianRidge	0.63	0.63	0.61	0.01
ElasticNet	0.53	0.53	429.78	0.02	LassoCV	0.63	0.63	0.61	0.14
SGDRegressor	0.53	0.53	430.62	0.01	HuberRegressor	0.63	0.63	0.61	0.03
ElasticNetCV	0.52	0.52	433.05	0.05	ElasticNetCV	0.63	0.63	0.61	0.07
GradientBoostingRegressor	0.52	0.52	435.07	0.44	LarsCV	0.63	0.63	0.61	0.02

TweedieRegressor	0.51	0.51	438.63	0.83	OrthogonalMatchingPursuitCV	0.63	0.63	0.61	0.01
RANSACRegressor	0.50	0.51	440.89	0.13	LinearSVR	0.63	0.63	0.61	0.07
HuberRegressor	0.50	0.50	442.22	0.05	SGDRegressor	0.62	0.62	0.62	0.02
PassiveAggressiveRegressor	0.50	0.50	443.12	0.02	OrthogonalMatchingPursuit	0.58	0.59	0.64	0.01
ExtraTreeRegressor	0.50	0.50	443.86	0.01	DecisionTreeRegressor	0.57	0.58	0.65	0.04
PoissonRegressor	0.48	0.48	450.43	5.12	AdaBoostRegressor	0.55	0.55	0.67	0.16
LinearSVR	0.45	0.45	464.58	0.01	TweedieRegressor	0.53	0.53	0.69	0.20
KernelRidge	0.40	0.40	484.34	0.40	ExtraTreeRegressor	0.48	0.48	0.72	0.03
OrthogonalMatchingPursuit	0.38	0.38	492.94	0.01	RANSACRegressor	0.40	0.41	0.77	0.07
BaggingRegressor	0.37	0.37	498.33	0.14	ElasticNet	0.25	0.25	0.87	0.03
DecisionTreeRegressor	0.35	0.35	504.85	0.03	Lasso	-0.01	-0.00	1.00	0.02
NuSVR	0.00	0.01	624.69	0.38	DummyRegressor	-0.01	-0.00	1.00	0.01
SVR	-0.00	0.00	626.32	0.55	LassoLars	-0.01	-0.00	1.00	0.02
DummyRegressor	-0.01	-0.00	627.76	0.01	PassiveAggressiveRegressor	-0.21	-0.20	1.10	0.01
QuantileRegressor	-0.07	-0.07	647.68	1008.94	GaussianProcessRegressor	-47.58	-47.27	6.96	0.81
AdaBoostRegressor	-0.40	-0.39	740.65	0.12					
GammaRegressor	-15.23	-15.13	2520.11	0.64					
GaussianProcessRegressor	-44.15	-43.85	4202.71	0.86					

Table S2. Results for Extra-Trees Regressor 5-fold cross-validation with no data transformation.

fit_time	score_time	test_r2	test_neg_mean_absolute_error
0.60	0.02	0.74	-92.49
0.59	0.02	0.87	-77.19
0.59	0.02	-2.91	-141.35
0.59	0.02	0.76	-88.99
0.58	0.02	0.40	-140.24

Table S3. Light Gradient-Boosting Model parameters.

Parameters	Grid	Random
boosting_type	gbdt	gbdt
class_weight	None	None
colsample_bytree	0.5	0.6603928648679768
importance_type	split	split
learning_rate	0.1	0.07252142853404991
max_depth	-1	6
min_child_samples	20	2
min_child_weight	0.001	0.001
min_split_gain	0.0	0.0
n_estimators	300	859
n_jobs	None	None
num_leaves	50	31
objective	None	None
random_state	None	None
reg_alpha	0.1	0.9121916727240686
reg_lambda	1.0	0.6313675599517621
subsample	1.0	1.0
subsample_for_bin	200000	200000
subsample_freq	0	0