Supporting Information for

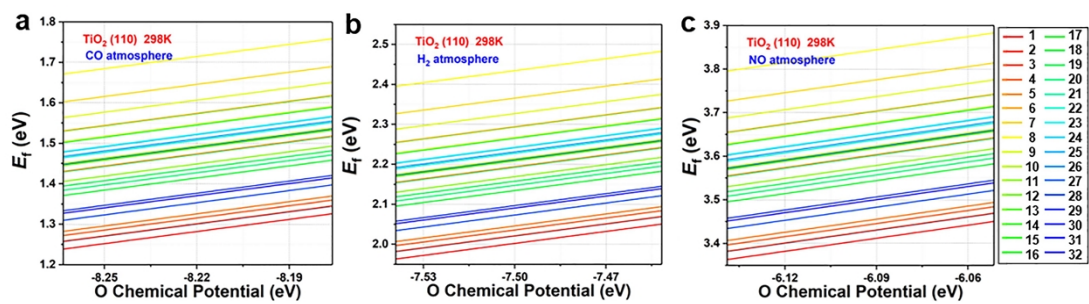**A machine learning-assisted study of the formation of oxygen**

vacancies in anatase titanium dioxide

Dan Wang,[+,a] Ronghua Zan, [+,b] Xiaorong Zhu,[*,c] Yuwei Zhang, [a] Yu Wang, [a] Yanhui Gu,[*,b] and Yafei Li[*,a]
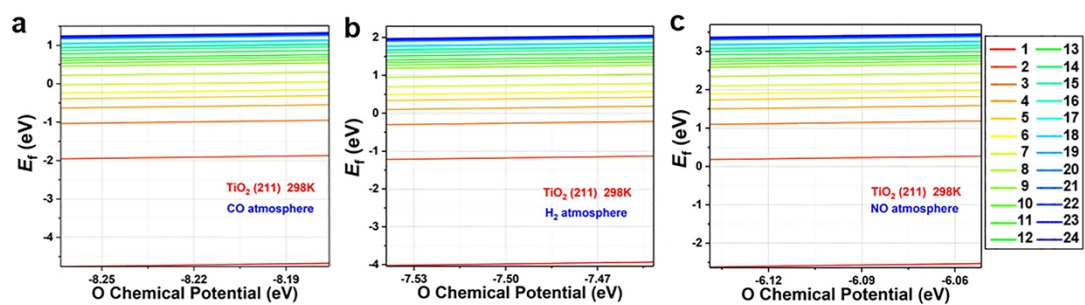
[a]Jiangsu Key Laboratory of New Power Batteries, Jiangsu Collaborative Innovation Centre of Biomedical Functional Materials, School of Chemistry and Materials Science, Nanjing Normal University, Nanjing, 210023, P. R. China

[b]School of Computer Science and Technology, Nanjing Normal University, Nanjing 210023, P. R. China
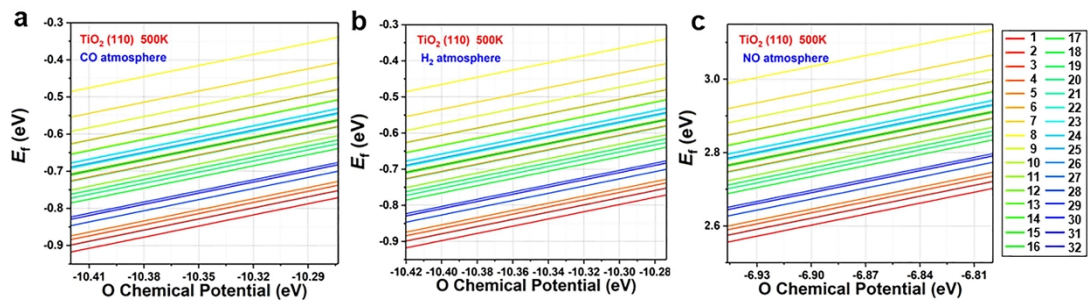
[c]College of Chemistry and Chemical Engineering, Nantong University, Nantong 226019, P. R. China
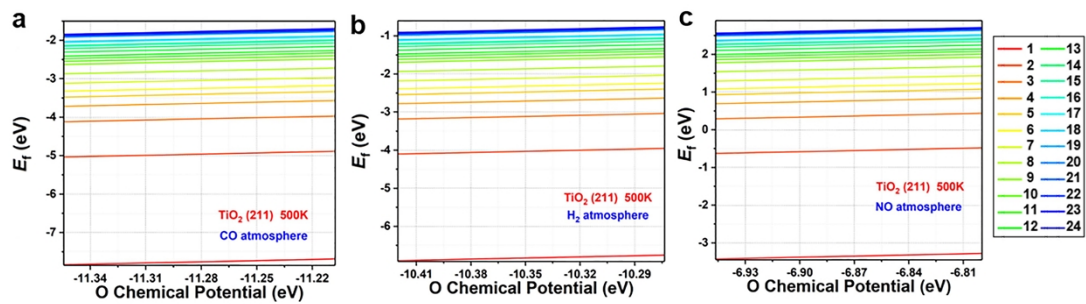
Figure S1. The O vacancy formation energy of the uppermost layer of O atoms on TiO$_2$ (110), under CO, H$_2$ and NO atmosphere when set the reaction temperature to be 298 K.

Figure S2. The O vacancy formation energy of the uppermost layer of O atoms on TiO$_2$ (211), under CO, H$_2$ and NO atmosphere when set the reaction temperature to be 298 K.

Figure S3. The O vacancy formation energy of the uppermost layer of O atoms on TiO$_2$ (110), under CO, H$_2$ and NO atmosphere when set the reaction temperature to be 500 K.
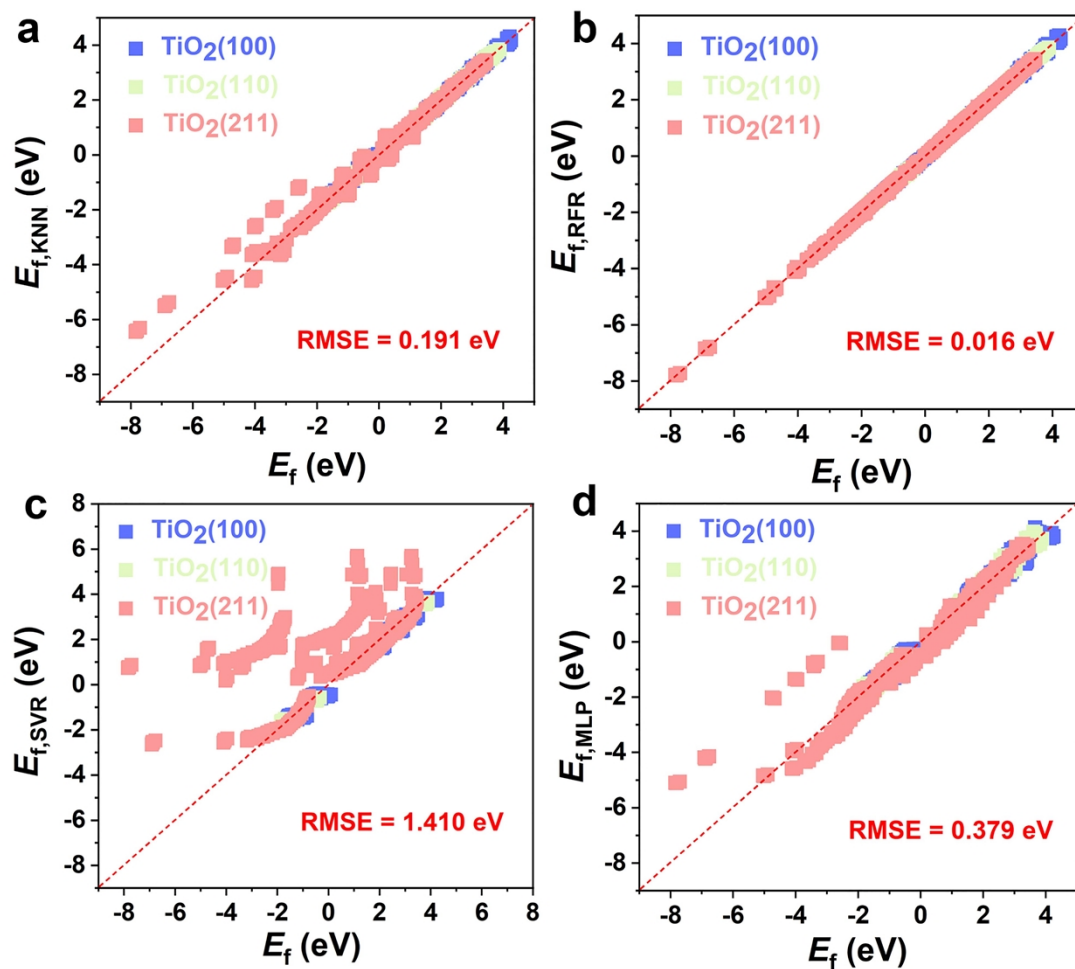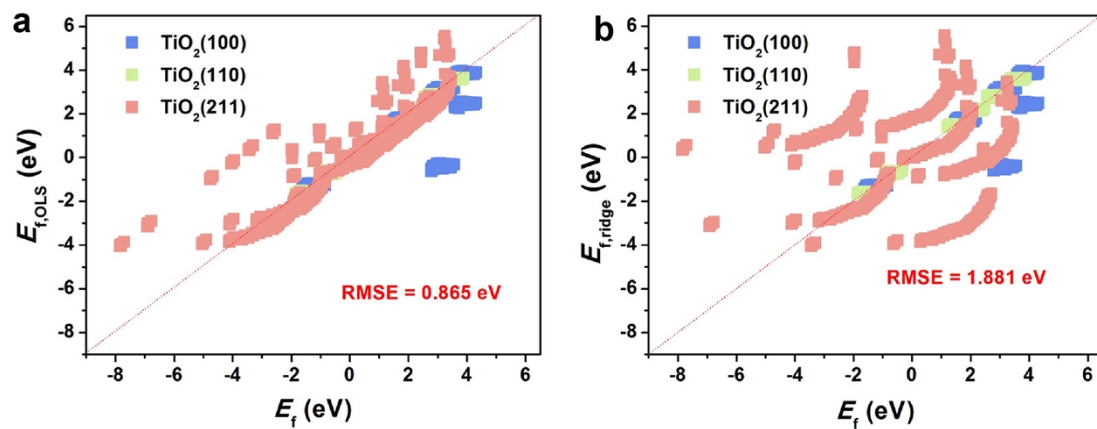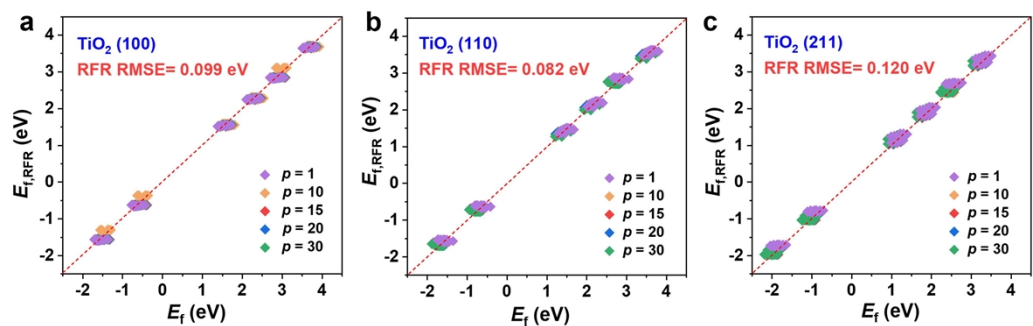
Figure S4. The O vacancy formation energy of the uppermost layer of O atoms on TiO$_2$ (211), under CO, H$_2$ and NO atmosphere when set the reaction temperature to be 500 K.

Figure S5. The compassion between DFT calculated E_f and (a)KNN, (b) RFR, (c)SVR, (d)MLP predicted values on TiO$_2$(100), TiO$_2$(110), TiO$_2$(211).

Figure S6. The compassion between DFT calculated E_f and (a)OLS, (b)ridge predicted values on $TiO_2(100)$, $TiO_2(110)$, $TiO_2(211)$.

Figure S7. Comparison between the E_f of the second O layer predicted by the ML based on the RFR model and the results calculated by DFT for (a) TiO$_2$(211), (b) TiO$_2$(110), and (c) TiO$_2$(100) surfaces. The purple, yellow, red, blue, and green dots represent the partial pressure ratio of 1, 10, 15, 20, and 30, respectively.

Table S1. The SISSO algorithm uses different numbers of features to construct equations for E_f.

Features	Equation	RMSE
$n, p, T, \gamma,$ μ_1, μ_2, μ_3	$E_f = (0.943(p+n) - 0.640(p \times \gamma))\left(\dfrac{\mu_3}{n}\right) - 0.301\dfrac{p \times \mu_3}{n \times \gamma} + 9.154$	0.7106 eV
$n, T, \gamma,$ μ_1, μ_2, μ_3	$E_f = \left(2.973\mu_3 - 2.049(\mu_3 \times \gamma) - 3.164\dfrac{\gamma}{n}\right)\gamma + 9.356$	0.7112 eV
$n, T, \gamma,$ μ_2, μ_3	$E_f = \left(2.973\mu_3 - 2.049(\mu_3 \times \gamma) - 3.164\dfrac{\gamma}{n}\right)\gamma + 9.356$	0.7112 eV
$n, \gamma, \mu_2,$ μ_3	$E_f = \left(2.973\mu_3 - 2.049(\mu_3 \times \gamma) - 3.164\dfrac{\gamma}{n}\right)\gamma + 9.356$	0.7112 eV
γ, μ_2, μ_3	$E_f = (0.035\mu_3 + 4.024 - 2.525\gamma)(\mu_3 \times \gamma) + 11.367$	0.8003 eV
γ, μ_3	$E_f = 7.832\mu_3 + 2597.384\dfrac{1}{\mu_3} + 8401.311\dfrac{1}{\mu_3^2} + 254.794$	1.0002 eV