Supplementary information

p-d Coupling: Prerequisite for Band-Like Doping Levels in Metal Oxides

Kangyu Zhang[1,2], Li-Chang Yin[1,2*], Guoqiang Deng[1,2], Xing-Qiu Chen[1,2], Hui-Ming Cheng[1,3], Gang Liu[1,2*]

[1] Shenyang National Laboratory for Materials Science, Institute of Metal Research, Chinese Academy of Sciences, 72 Wenhua Road, Shenyang 110016, China.

[2] School of Materials Science and Engineering, University of Science and Technology of China, 72 Wenhua Road, Shenyang 110016, China.

[3] Institute of Technology for Carbon Neutrality, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, 1068 Xueyuan Blvd, Shenzhen 518055, China.

*Correspondence: lcyin@imr.ac.cn (LCY); gangliu@imr.ac.cn (GL)

Structure descriptor construction and XGBoost model training

In order to apply ML methods for predicting the bandgap values of all B/N-codoping configurations, each B/N-codoping configuration has to be represented by fixed-length structure descriptor that accurately encodes the relationship between dopant spatial orderings with the corresponding bandgap value without the need of DFT calculations. Considering that B/N-codoping essentially induces charge redistribution between the B/N dopants in anatase TiO_2, and the features of charge redistribution, as determined by the B/N dopant spatial orderings, yield doping levels of different widths and energies. Therefore, accurate structure descriptors can be constructed by characterizing the charge redistribution between the B/N dopants in anatase TiO_2.

As shown in our previous work[1] of predicting electronic properties of N and oxygen vacancy (V_O) codoped TiO_2, the charge redistribution between two dopants can be accurately characterized by the shortest sequences consisting of adjacent Ti-O/N atoms that connect the two dopants, termed charge transfer path. It should be noted that, the charge transfer path is permutation invariant due to its sole dependence on the bonding patterns of TiO_2, regardless of atomic indexing, this ensures that the constructed structure descriptor is also permutation invariant. Therefore, we also applied this concept of charge transfer path for characterizing the charge redistribution between the B/N dopants in this work. For instance, there are in total two possible charge transfer paths connecting the N dopant (N1) and one of the four adjacent Ti atoms of interstitial B (Ti1) as depicted in Fig. S1. For each possible charge transfer path, the occurrences of all types of triatomic fragments, such as Ti-O-Ti@102°, were separately counted to construct a 24×1 array. Subsequently, by adding the two arrays, the charge redistribution between the Ti1 and N1 can be characterized by a 24×1 array denoted as Ti1-N1 in Fig. S1.

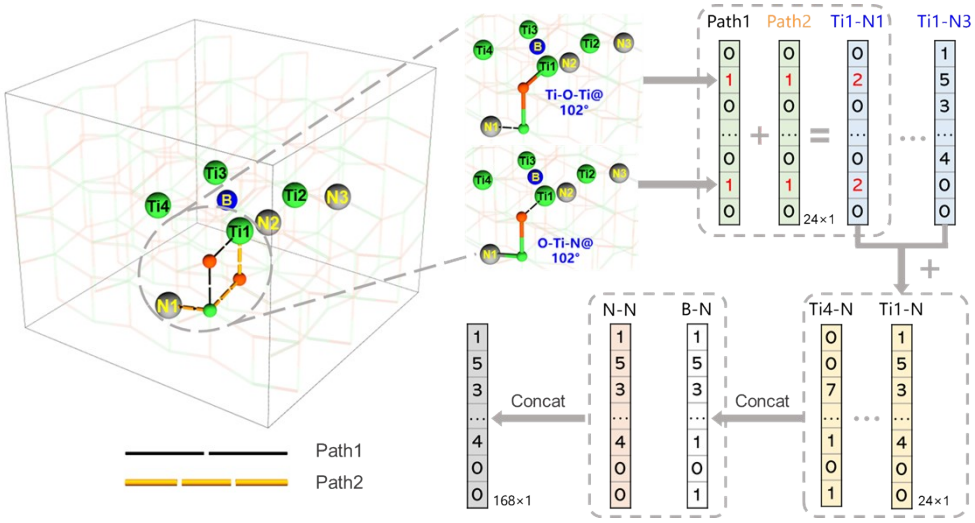

Fig. S1 Schematic illustration of structure descriptor construction for a B/N-codoping configuration in anatase TiO_2.

Furthermore, similar arrays for characterizing the charge redistribution between the Ti1 and the other two N dopants (N2 and N3) can be constructed, and the summation (denoted as Ti1-N in Fig. S1) fully characterizes the charge redistribution between the N dopants and the Ti1 atom. Similar to the Ti1-N array, three more arrays for the rest of the adjacent Ti atoms (Ti2, Ti3, and Ti4) of interstitial B can be respectively constructed, and concatenating these arrays with the Ti1-N array results in a 96×1 array denoted as B-N. As the B-N array naturally encodes the ease of achieving charge redistribution between B/N dopants via different adjacent Ti atoms, it effectively resolves the issue that the actual coordination of the interstitial B is unknown when constructing the structure descriptor for a B/N-codoping configuration. Besides the B-N array, a 72×1 array denoted as N-N, which is a concatenation of three 24×1 arrays, each characterizing the charge redistributions between a pair of N dopants, was also constructed. By concatenating the 72×1 N-N array with the 96×1 B-N array, a 168×1 array was constructed for representing a B/N-codoping configuration in anatase TiO_2 as depicted in Fig. S1.

Besides making the B/N-codoping configurations compatible with ML methods, the constructed structure descriptors were also applied to ensure the representativeness of samples in the training set, *i.e.*, a small set of B/N-codoping configurations for DFT calculations. Specifically, we first compute vector similarities between all pairs of structure descriptors. Starting from the structure descriptor with the highest similarity to the rest, we iteratively find the most similar structure descriptor to the previous one, thereby sorting all structure descriptors based on vector similarities. After sorting, every $5^{th}/20^{th}$ structure descriptor was sampled from all the 8144 ones, which corresponds to a set of 915 B/N-codoping configurations whereby the similarities among them are the lowest. Subsequently, we split the 915 samples into a training set (80%), a validation set (10%), and a testing set (10%) for model evaluation. Based on these datasets, a highly accurate XGBoost model for predicting the bandgap values of B/N-codoping configurations was trained, as summarized in Fig. S2.

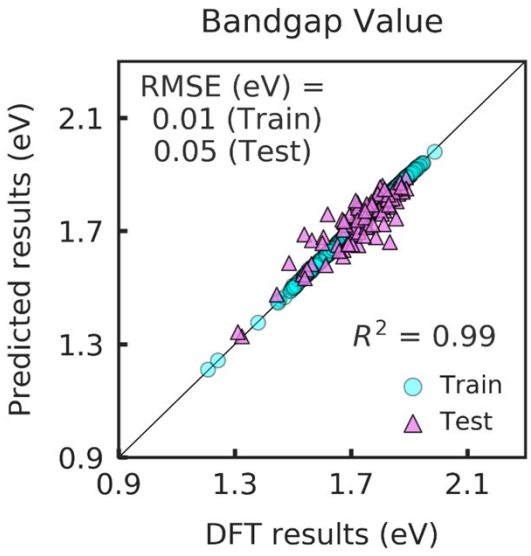

Fig. S2 Parity plot of bandgap values of B/N-codoping configurations in anatase TiO_2. The x-axis corresponds to the DFT calculated bandgap values of B/N-codoping configurations, and the y-axis corresponds to the respective bandgap values predicted by the trained XGBoost model.

Moreover, in order to demonstrate the high accuracy of our newly proposed structure descriptor in predicting the bandgap values of B/N-codoping configurations, in comparison to other methods, we also trained several crystal graph convolutional neural network (CGCNN) models[2], as well as XGBoost models using the Dexp2 structure descriptor proposed by Kaneko et al.[3], for predicting the bandgap values of B/N-codoping configurations on the same dataset in this work.

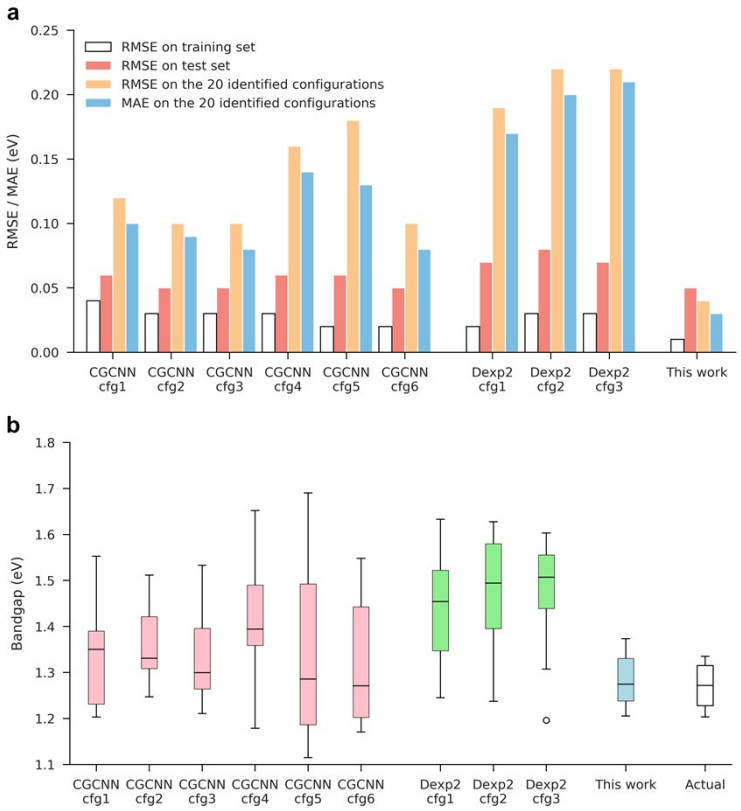

Fig. S3 Performance comparisons of the newly proposed structure descriptor with other methods. **a,** Performances of ML models in predicting the bandgap values of B/N-codoping configurations in anatase TiO_2. The detailed model configurations are summarized in Table S5. **b,** Box plots of the bandgap values of the 20 identified B/N-codoping configurations in anatase TiO_2. Each box extends from the first quartile to the third quartile of the data, with a line at the median.

From the model performances summarized in Fig. S3a, among all considered models, the XGBoost model trained in this work using the newly proposed structure descriptors exhibits the highest accuracy. Additionally, even though only two of the 20 identified B/N-codoping configurations were included in the training dataset, the distribution of bandgap values of the 20 identified B/N-codoping configurations can be accurately reproduced by the XGBoost model trained using the newly proposed structure descriptor, as shown in Fig. S3b, thereby further demonstrating the high accuracy of our newly proposed structure descriptor in representing doping configurations.

DFT calculations

The bandgap values of the sampled 915 B/N-codoping configurations in anatase TiO_2 were calculated after structure relaxation by using the VASP code[4-7] with the projector augmented wave

method[8,9]. Considering that B/N-codoping barely affects the intrinsic band-edge of TiO_2, except that it induces N doping levels within the band gap, and the aim of predicting bandgap values using ML methods is to identify the B/N-codoping configurations that exhibit the largest bandgap narrowing. Therefore, the exchange-correlation term was described by the Perdew-Burke-Ernzerhof (PBE) functional[10]. The cutoff energy of plane-wave basis was set to 500 eV, and Gamma point was used for sampling the Brillouin zone in geometry optimizations, where both the cell parameters and atomic positions of the B/N-codoping configurations were optimized until the force on each ion was smaller than 0.05 eV/Å.

Besides anatase TiO_2, the DFT calculations for rutile and brookite TiO_2 were performed using the same setup, except that the cell parameters and atomic positions were optimized until the force on each ion was smaller than 0.01 eV/Å. As well known, the PBE functional tends to produce delocalized electronic states due to the excessive repulsion caused by the so-called self-interaction error (SIE)[11]. Therefore, we also applied the HSE06 functional[12-14] to double check the electronic structures of a few specific B/N-codoping configurations in TiO_2. Specifically, the fraction of exact exchange in HSE06 functional is set to 0.2 $Å^{-1}$ for anatase and rutile TiO_2, and 0.15 $Å^{-1}$ for brookite TiO_2. With these settings, the bandgap values of pristine anatase, rutile and brookite TiO_2 are calculated to be 3.3 eV, 3.0 eV and 3.3 eV, respectively.

It should be noted that, the interstitial B in anatase TiO_2 has three different types of coordinations[15], as depicted in Fig. S4. In order to specify a protocol to deal with interstitial B when performing high-throughput DFT calculations for obtaining bandgap values, we first scrutinized the effect of the coordinations of interstitial B on the electronic structures of B/N-codoping configurations. Specifically, we selected 117 B/N-codoping configurations from the 915 ones and replaced the four-coordinated B in each with the other two types of three-coordinated B, resulting in six new B/N-codoping configurations per sample. These modified B/N-codoping configurations were then compared based on DFT calculations. As summarized in Table S4, for a fixed N dopant spatial ordering, the three-coordinated B in basal plane is energetically the most favorable, which is in accordance with the previous study[15]. Most importantly, the bandgap value is barely affected by specific coordination of interstitial B. Therefore, it is reasonable to consider only four-coordinated B in high-throughput DFT calculations for the purpose of screening B/N-codoping configurations with the most significant bandgap narrowing.

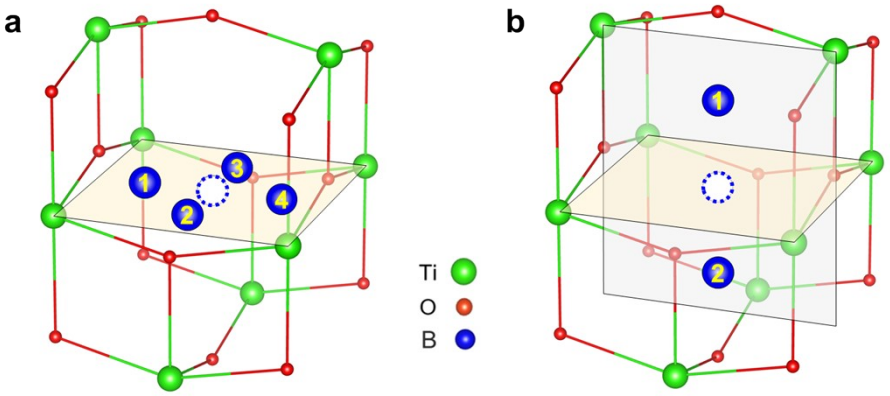

Fig. S4 Different types of interstitial B in anatase TiO_2. a Four possible positions of a three-coordinated B in the basal plane of interstitial cavity. **b** Two possible positions of a three-coordinated B in the plane normal to the basal plane of interstitial cavity. The four-coordinated B in interstitial cavity is illustrated by a dotted circle in both **a** and **b**.

Besides TiO_2, the electronic structures of doping configurations in $ScTaO_4$ were calculated by using PBE functional since it can well reproduce the experimental bandgap of pristine $ScTaO_4$ (experimental 3.75 eV[16], calculated 4.03 eV). As for the other three metal oxides, the DFT + *U* method[17] was applied to describe the electronic structures of the few specific doping configurations that exhibit uniform N dopant spatial orderings. Specifically, a *U* value of 5 eV was applied to the metal *d* orbitals in WO_3 and SnO_2. As for $MgTa_2O_6$, a *U* value of 5 eV was applied to Ta 5*d* orbitals, and a *U* value of 8 eV was applied to 2*p* orbitals in O and N elements by referring to previous theoretical studies[18,19].

References

1. Zhang, K. Y., Yin, L. C., Liu, G. & Cheng, H. M. Accurate structural descriptor enabled screening for nitrogen and oxygen vacancy codoped TiO_2 with a large bandgap narrowing. *J. Mater. Sci. Technol.* **122**, 84-90 (2022).
2. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
3. Kaneko, M., Fujii, M., Hisatomi, T., Yamashita, K. & Domen, K. Regression model for stabilization energies associated with anion ordering in perovskite-type oxynitrides. *J. Energy Chem.* **36**, 7-14 (2019).
4. Kresse, G. & Hafner, J. Ab initio molecular-dynamics for liquid-metals. *Phys. Rev. B* **47**, 558-561 (1993).

5. Kresse, G. & Hafner, J. Ab-initio molecular-dynamics simulation of the liquid-metal amorphous-semiconductor transition in germanium. *Phys. Rev. B* **49**, 14251-14269 (1994).
6. Kresse, G. & Furthmuller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15-50 (1996).
7. Kresse, G. & Furthmuller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169-11186 (1996).
8. Blochl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953-17979 (1994).
9. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758-1775 (1999).
10. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865-3868 (1996).
11. Cohen, A. J., Mori-Sanchez, P. & Yang, W. T. Insights Into Current Limitations of Density Functional Theory. *Science* **321**, 792-794 (2008).
12. Heyd, J. & Scuseria, G. E. Efficient Hybrid Density Functional Calculations in Solids: Assessment of the Heyd-Scuseria-Ernzerhof Screened Coulomb Hybrid Functional. *J. Chem. Phys.* **121**, 1187-1192 (2004).
13. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid Functionals based on a Screened Coulomb Potential. *J. Chem. Phys.* **118**, 8207-8215 (2003).
14. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Erratum: “Hybrid Functionals based on a Screened Coulomb Potential” [J. Chem. Phys. 118, 8207 (2003)]. *J. Chem. Phys.* **124**, 219906 (2006).
15. Finazzi, E., Di Valentin, C. & Pacchioni, G. Boron-doped anatase TiO_2: Pure and hybrid DFT calculations. *J. Phys. Chem. C* **113**, 220-228 (2009).
16. Pei, L. *et al.* A novel visible-light-responsive semiconductor $ScTaO_{4-x}N_x$ for photocatalytic oxygen and hydrogen evolution reactions. *ChemCatChem* **13**, 180-184 (2021).
17. Dudarev, S. L., Botton, G. A., Savrasov, S. Y., Humphreys, C. J. & Sutton, A. P. Electron-energy-loss spectra and the structural stability of nickel oxide: An LSDA+U study. *Phys. Rev. B* **57**, 1505-1509 (1998).
18. Kim, J. Y. *et al.* Electronic structure and stability of low symmetry Ta_2O_5 polymorphs. *Phys. Status Solidi-Rapid Res. Lett.* **8**, 560-565 (2014).
19. Droghetti, A., Pemmaraju, C. D. & Sanvito, S. Polaronic distortion and vacancy-induced magnetism in MgO. *Phys. Rev. B* **81**, 092403 (2010).

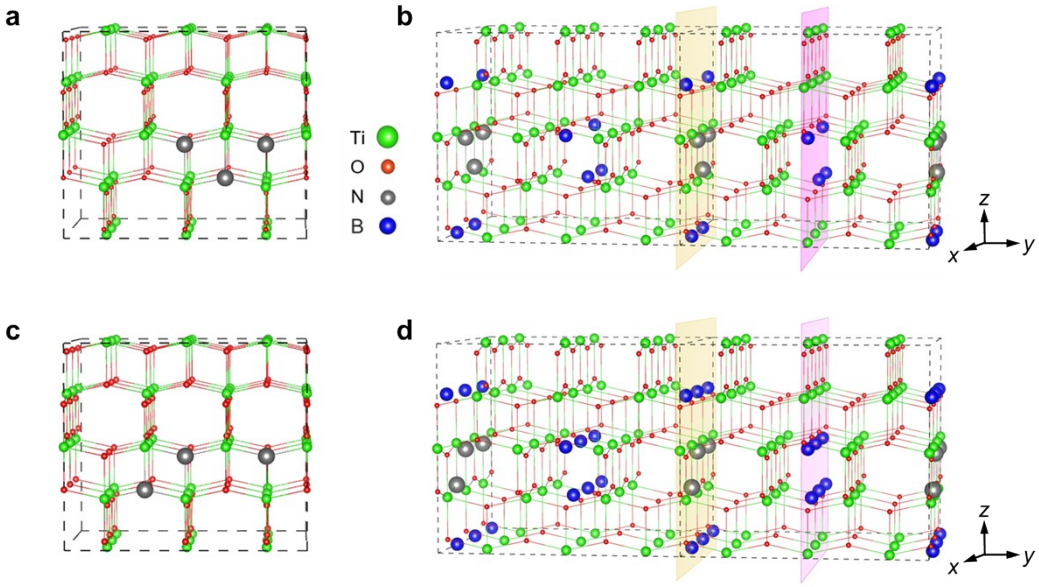

Fig. S5 Atomic structures of the identified 20 B/N-codoping configurations. a and **b,** Illustrations of the first type of characteristic N dopant spatial ordering and all 8 possible positions of interstitial B. **c** and **d,** Illustrations of the second type of characteristic N dopant spatial ordering and all 12 possible positions of interstitial B.

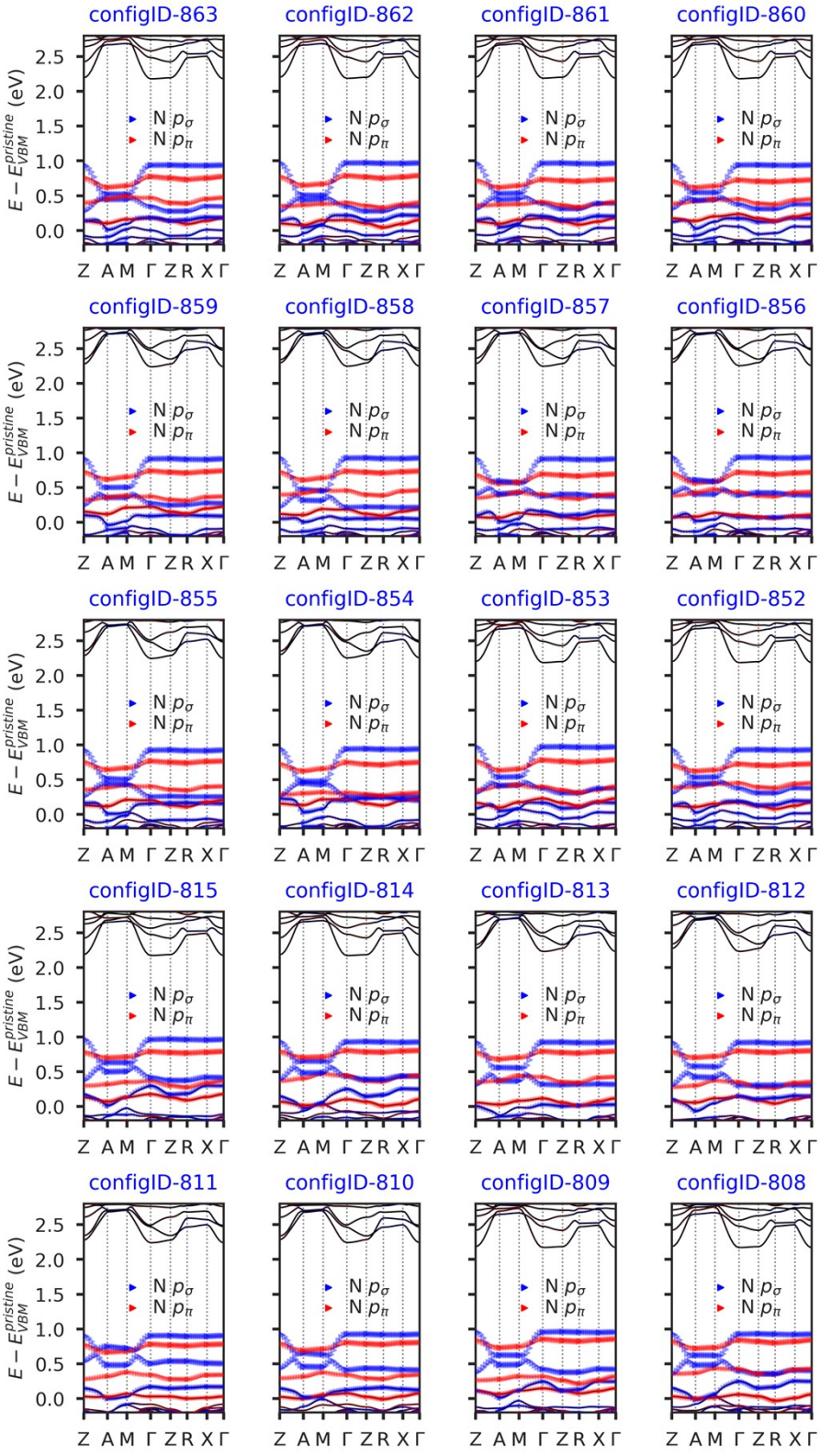

Fig. S6 Band structures of the identified 20 B/N-codoping configurations in anatase TiO_2. The energy values are referenced to the VBM of pristine anatase TiO_2, which is set to be 0.0 eV.

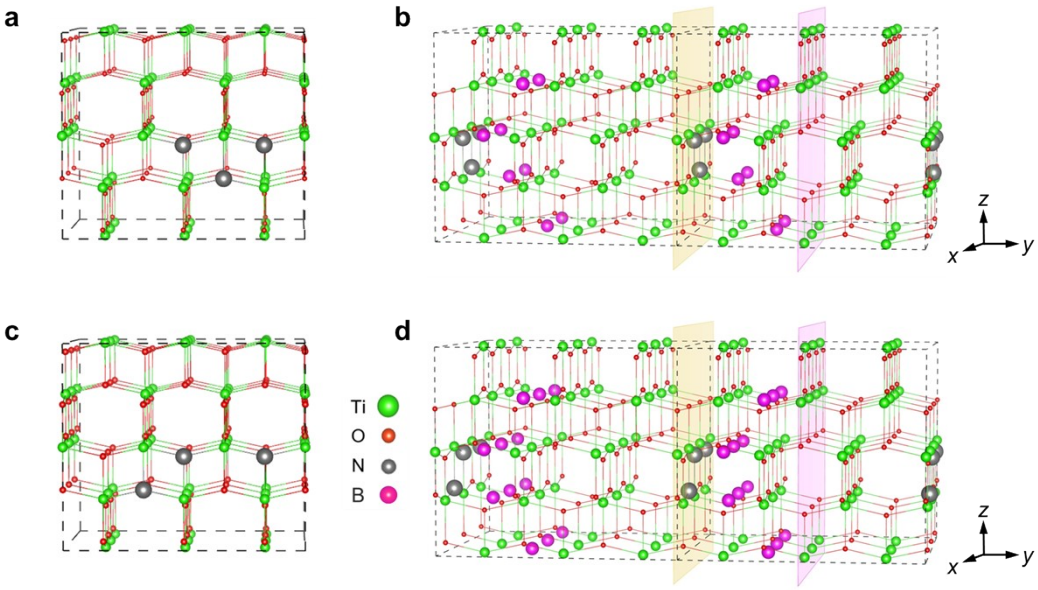

Fig. S7 Atomic structures of the 20 B/N-codoping configurations that exhibit the characteristic N dopant spatial orderings while the interstitial B is located elsewhere. a and **b,** Illustrations of the first type of characteristic N dopant spatial ordering and all 8 possible positions of interstitial B not on the blue/purple (010) planes. **c** and **d,** Illustrations of the second type of characteristic N dopant spatial ordering and all 12 possible positions of interstitial B not on the blue/purple (010) planes.

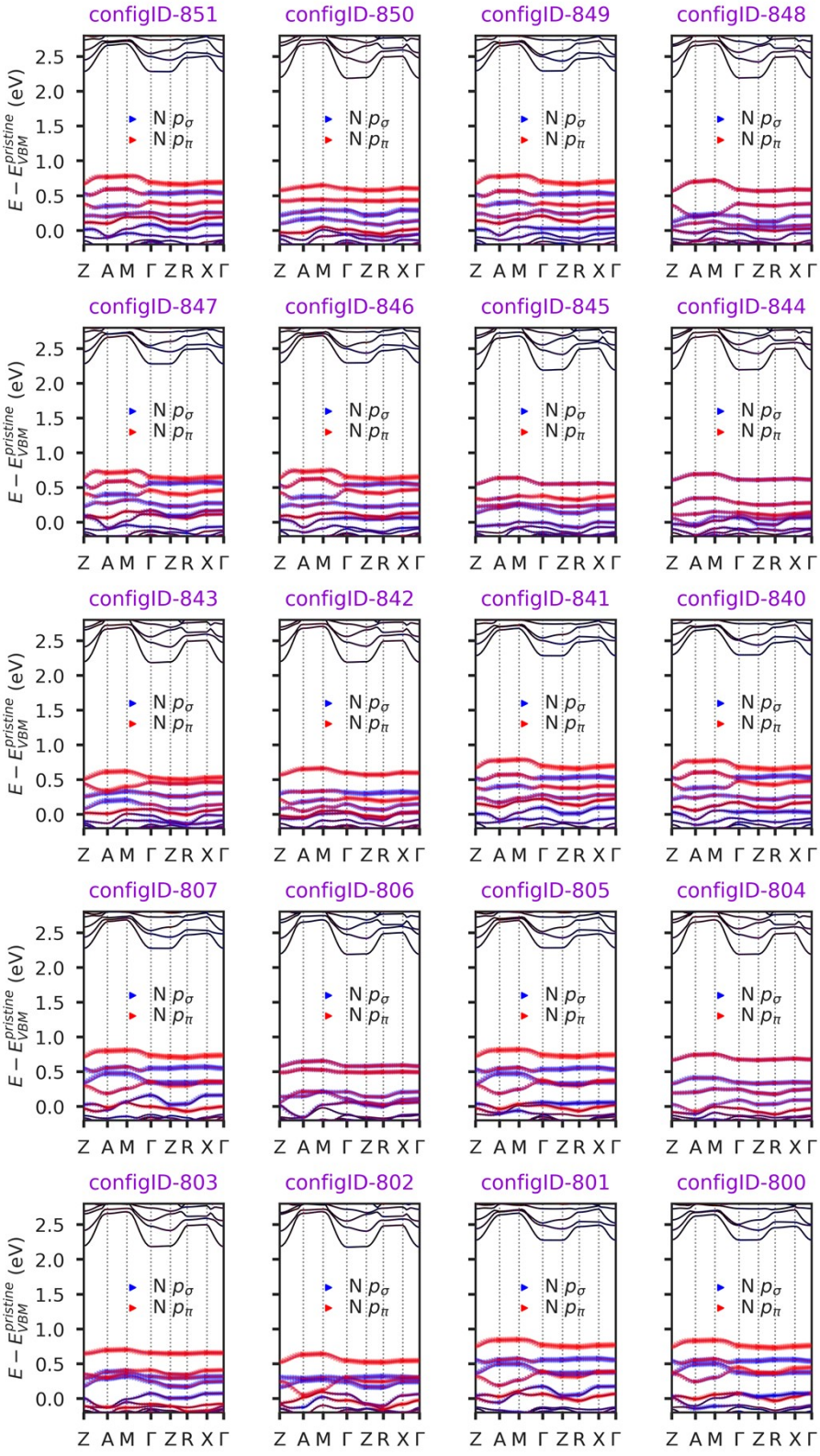

Fig. S8 Band structures of all the 20 B/N-codoping configurations that exhibit the characteristic N dopant spatial orderings while the interstitial B is located elsewhere. The energy values are referenced to the VBM of pristine anatase TiO_2, which is set to be 0.0 eV.

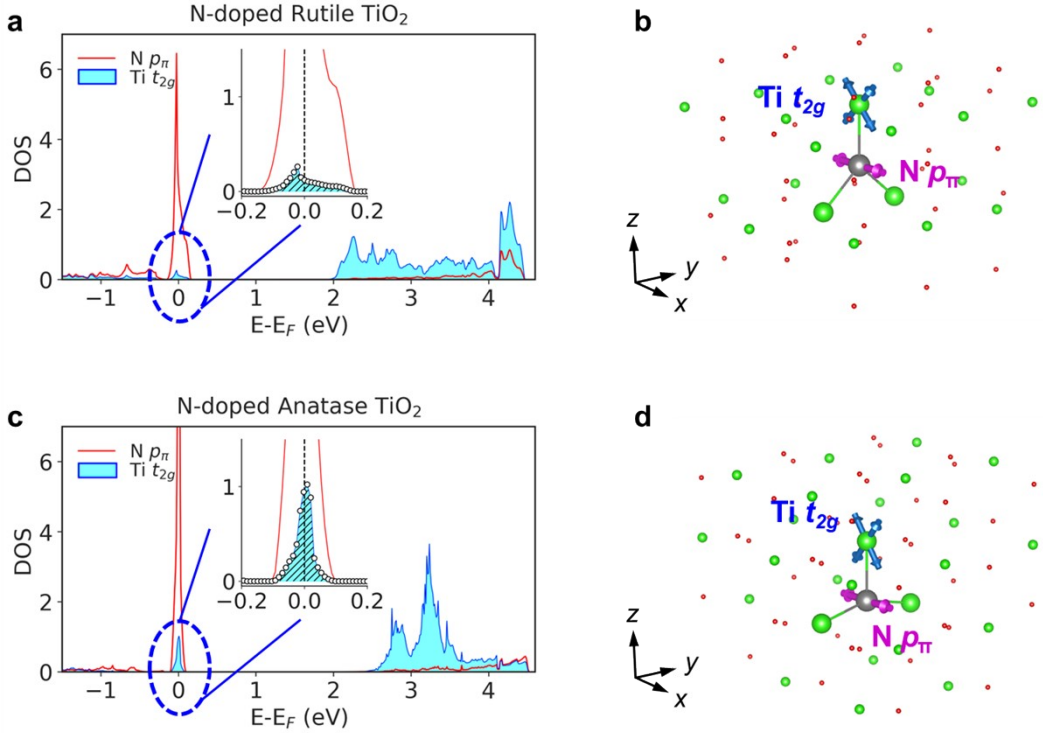

Fig. S9 Comparison of *p-d* coupling in rutile and anatase TiO_2. a, Partial density of states (PDOS) of N-p_π and Ti-t_{2g} orbitals in N-doped rutile TiO_2. **b,** The illustrations of N-p_π and Ti-t_{2g} orbitals in rutile TiO_2. **c,** PDOS of N-p_π and Ti-t_{2g} orbitals in N-doped anatase TiO_2. **d,** The illustrations of N-p_π and Ti-t_{2g} orbitals in anatase TiO_2.

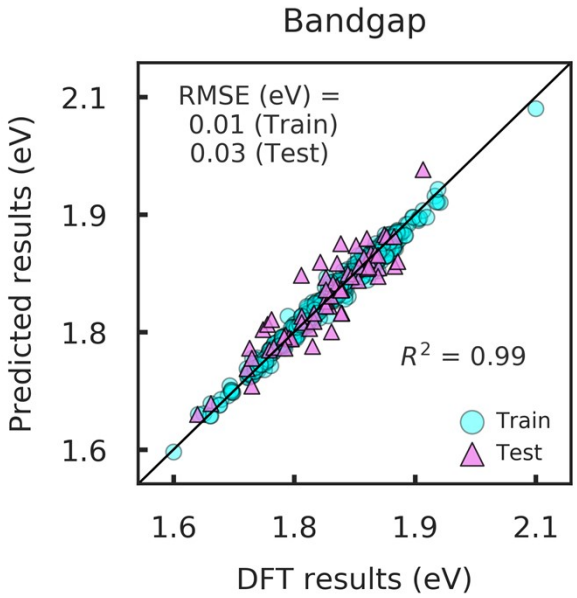

Fig. S10 Parity plot of the bandgap values of B/N-codoping configurations in rutile TiO_2. The x-axis corresponds to the DFT calculated bandgap values of B/N-codoping configurations, and the y-axis corresponds to the respective bandgap values predicted by the trained XGBoost model.

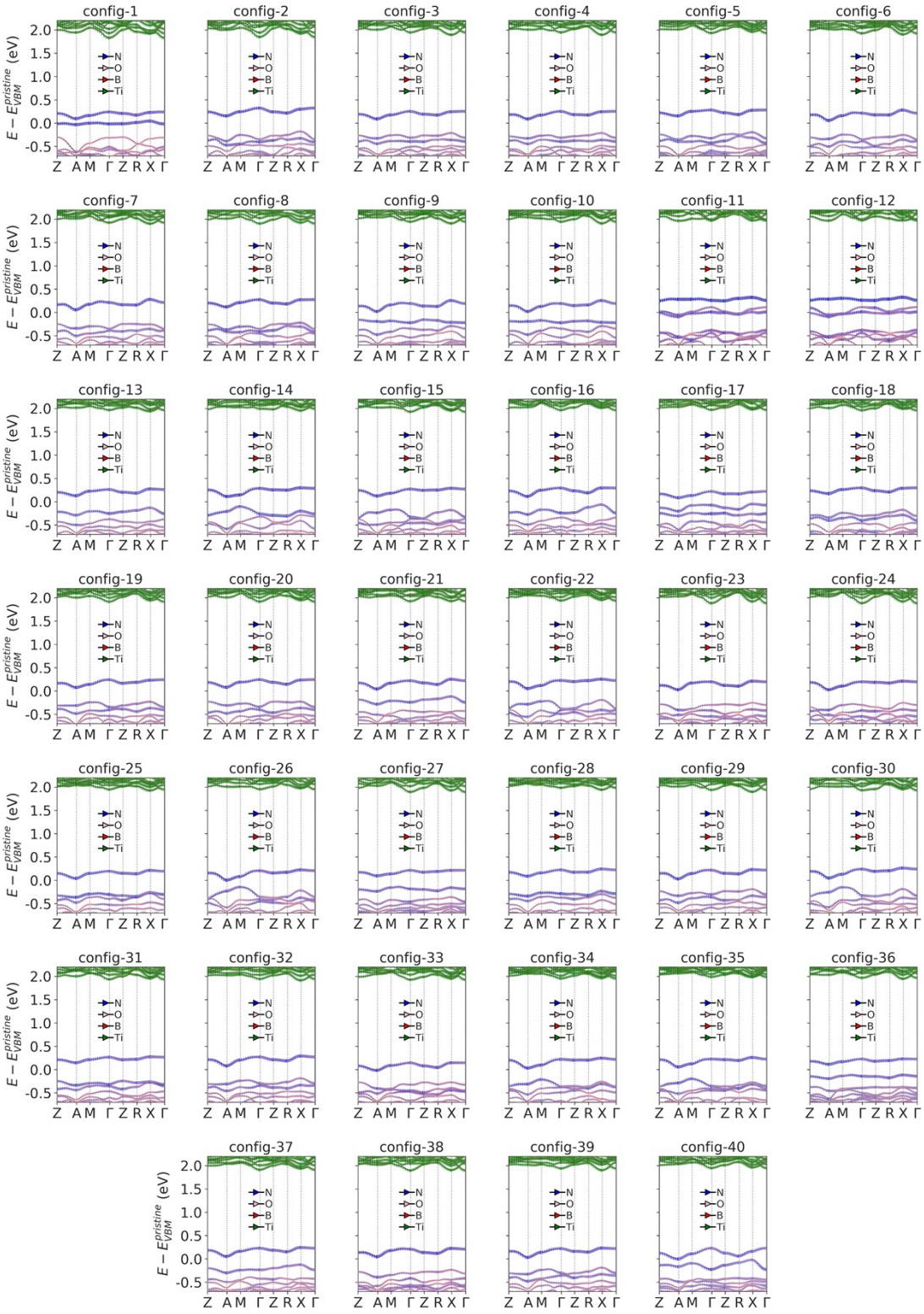

Fig. S11 Band structures of the predicted 40 B/N-codoping configurations with the most significant bandgap narrowing in rutile TiO_2. The energy values are referenced to the VBM of pristine rutile TiO_2, which is set to be 0.0 eV.

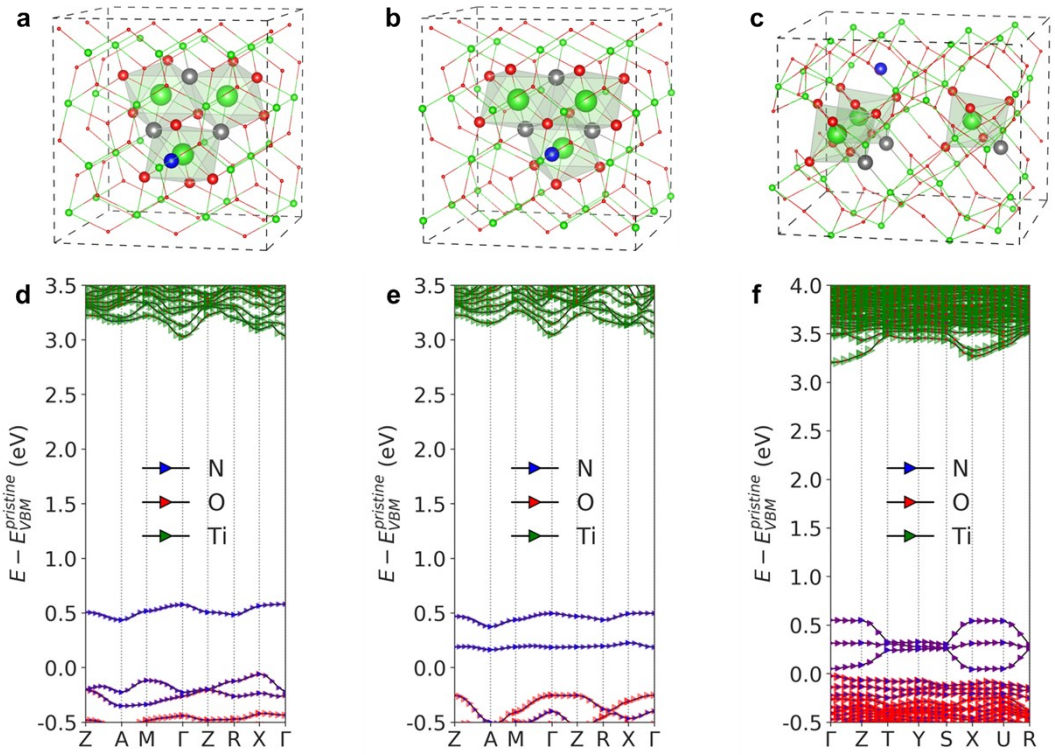

Fig. S12 Atomic structures and the corresponding electronic structures of B/N-codoping configurations in rutile and brookite TiO_2. a and **b,** Atomic structures of the two B/N-codoping configurations in rutile TiO_2 with the most significant bandgap narrowing. **c,** Atomic structure of a uniform B/N-codoping configuration in brookite TiO_2. **d-f,** Band structures at the HSE06 level of the B/N-codoping configurations illustrated in **a-c**, respectively. The energy values are referenced to the VBM of pristine rutile/brookite TiO_2, which is set to be 0.0 eV.

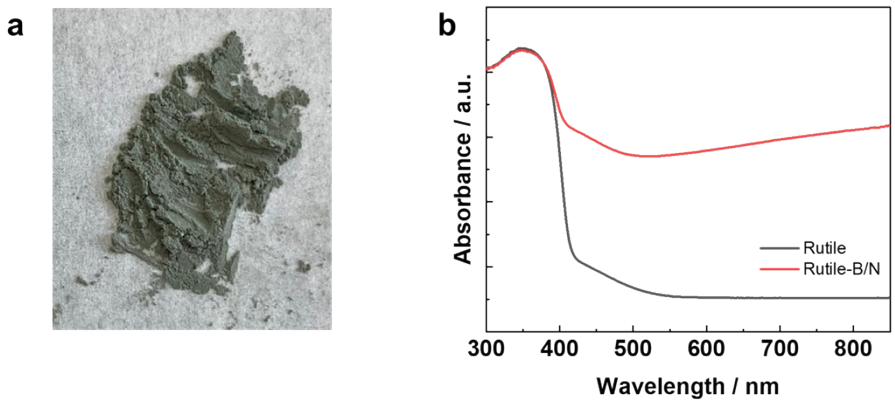

Fig. S13 Optical photograph and UV-visible absorption spectra of B/N-codoped rutile TiO_2. **a**, Optical photograph of the prepared B/N-codoped rutile TiO_2 sample. **b,** UV-visible absorption spectra of pristine and B/N-codoped rutile TiO_2.

Preparation of B/N-codoped rutile TiO_2

An amount of 75 mg of TiB_2 powder was suspended in a mixture solution of 8 mL of HCl aqueous solution (1.5 M) and 12 mL of isopropyl alcohol containing 25 mg of $NaNO_3$. The suspension was transferred to a Teflon-lined autoclave. The acidic hydrolysis of TiB_2 at 180 °C for 24 h in an autoclave led to the formation of rutile TiO_2 microspheres with the boron in the inner part. The product (B-doped rutile TiO_2) was collected by centrifugation, washed with deionized water three times to remove dissoluble ionic impurities, and dried at 100 °C. To prepare the B/N-codoped rutile TiO_2, the B-doped rutile TiO_2 was heated at 500 ^oC for 2 h with the NH_3 flow rate of 50 mL/min.

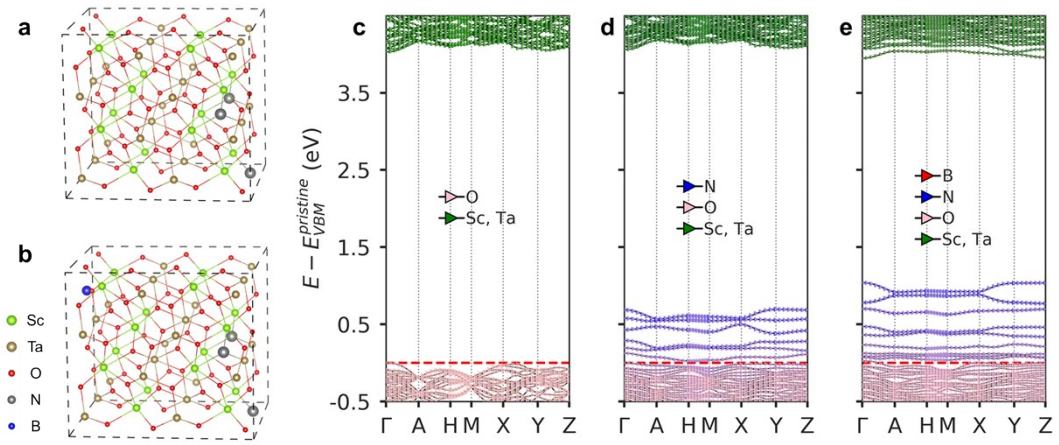

Fig. S14 The effect of uniform N-doping on the band structure of $ScTaO_4$. a and **b,** Atomic structures of a uniform N-doping configuration and a uniform B/N-codoping configuration in $ScTaO_4$. **c,** Band structure of pristine $ScTaO_4$. **d** and **e,** Band structures of the doping configurations illustrated in **a** and **b**, respectively. The energy values are referenced to the VBM of pristine $ScTaO_4$, which is set to be 0.0 eV.

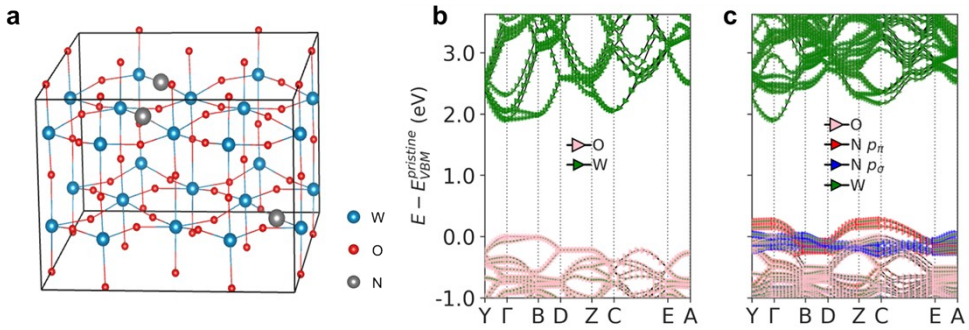

Fig. S15 The effect of uniform N-doping on the band structure of WO_3. a, Atomic structure of a uniform N doping configuration in WO_3. **b** and **c,** Band structures of pristine WO_3 and the uniform N doping configuration illustrated in **a**. The energy values are referenced to the VBM of pristine WO_3, which is set to be 0.0 eV.

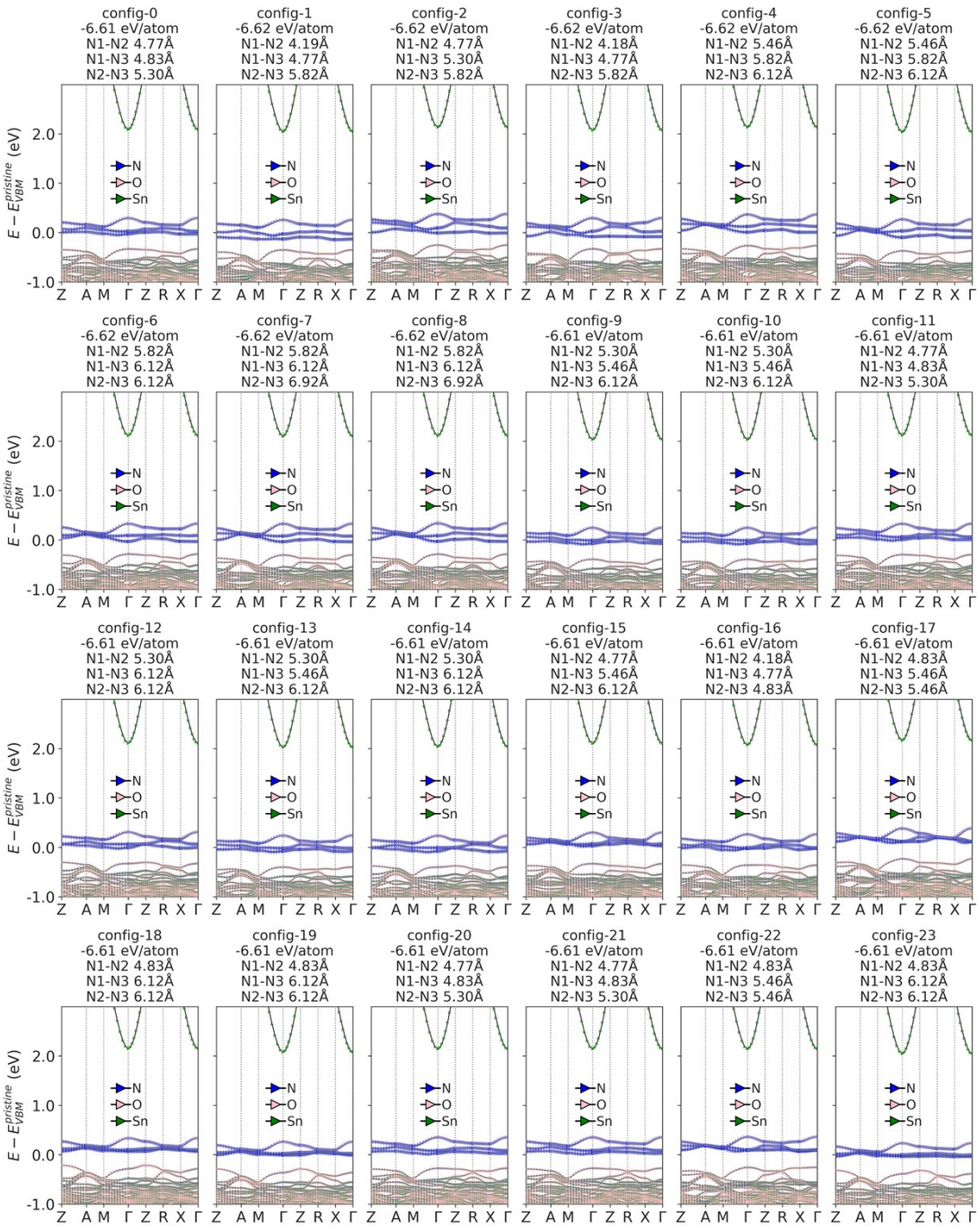

Fig. S16 Band structures of the 24 N-doping configurations in SnO_2 that exhibit the most uniform N dopant spatial orderings. The mutual distances between N dopants are annotated in the titles of subplots. The energy values are referenced to the VBM of pristine SnO_2, which is set to be 0.0 eV.

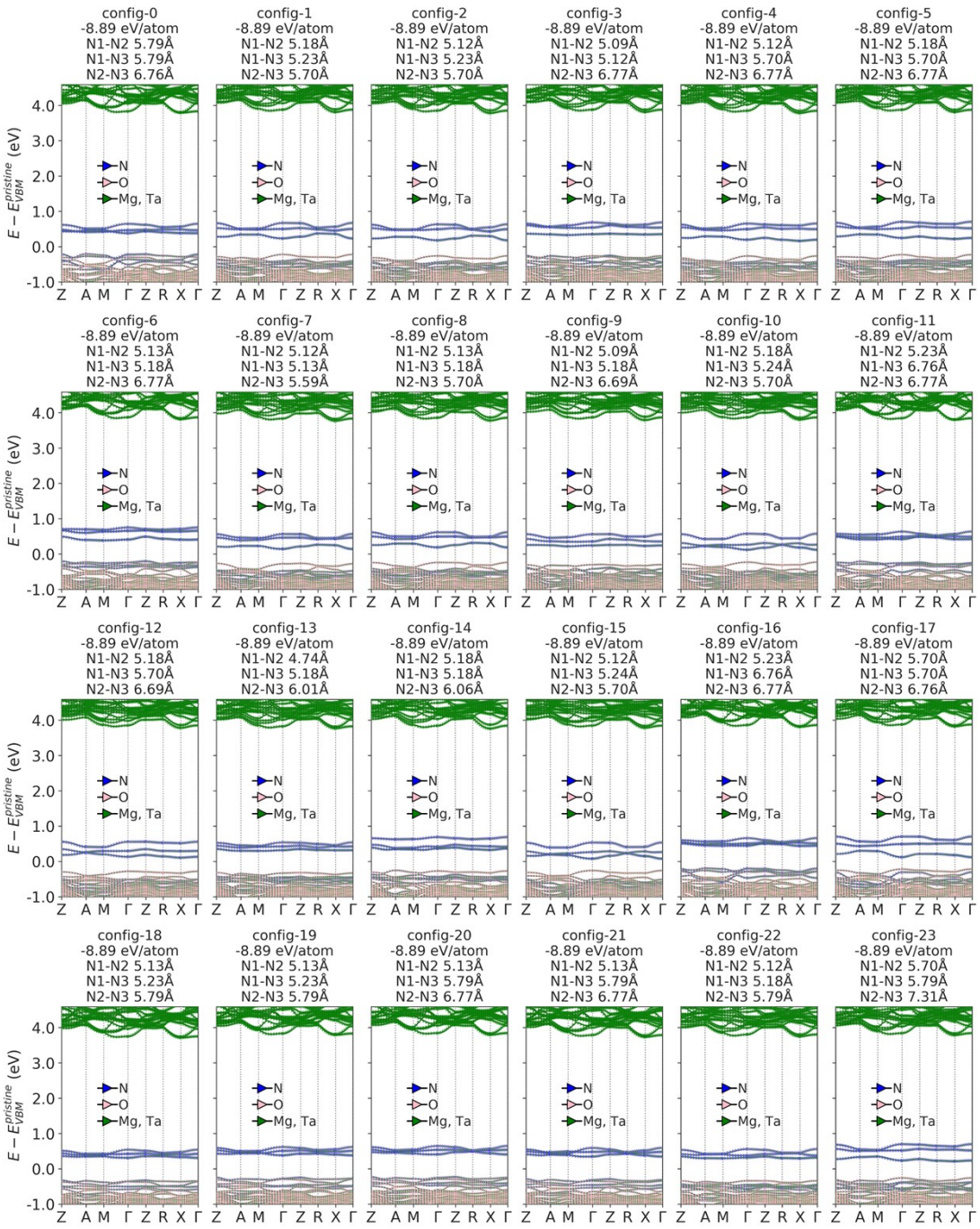

Fig. S17 Band structures of the 24 N-doping configurations in $MgTa_2O_6$ that exhibit the most uniform N dopant spatial orderings. The mutual distances between N dopants are annotated in the titles of subplots. The energy values are referenced to the VBM of pristine $MgTa_2O_6$, which is set to be 0.0 eV.

Table S1 Point groups of high-symmetry *k*-points in the first Brillouin zone of anatase, rutile and brookite TiO_2.

K point	Coordinate	Point group
Anatase TiO_2		
Z	(0.0,0.0,0.5)	C_{4v}
A	(0.5,0.5,0.5)	D_{2d}
M	(0.5,0.5,0.0)	D_{2h}
R	(0.0,0.5,0.5)	C_{2h}
X	(0.0,0.5,0.0)	C_{2v}
Γ	(0.0,0.0,0.0)	D_{4h}
Rutile TiO_2		
Γ	(0.0, 0.0, 0.0)	D_{4h}
Z	(0.0, 0.0, 0.5)	D_{4h}
A	(0.5, 0.5, 0.5)	D_{4h}
M	(0.5, 0.5, 0.0)	D_{4h}
R	(0.0, 0.5, 0.5)	D_{2h}
X	(0.0, 0.5, 0.0)	D_{2h}
Brookite TiO_2		
Γ	(0.0, 0.0, 0.0)	D_{2h}
R	(0.5, 0.5, 0.5)	D_{2h}
S	(0.5, 0.5, 0.0)	C_{2v}
T	(0.0, 0.5, 0.5)	D_{2h}
U	(0.5, 0.0, 0.5)	D_{2h}
X	(0.5, 0.0, 0.0)	D_{2h}
Y	(0.0, 0.5, 0.0)	D_{2h}
Z	(0.0, 0.0, 0.5)	D_{2h}

The high-symmetry *k*-point which allows *p*-*d* coupling is highlighted in blue.

Table S2 Point groups of high-symmetry *k*-points in the first Brillouin zone of $ScTaO_4$, WO_3, SnO_2 and $MgTa_2O_6$.

K point	Coordinate	Point group
$ScTaO_4$		
Γ	(0.0, 0.0, 0.0)	C_{2h}
A	(-0.5, 0.5, 0.0)	C_{2h}
C	(-0.5, 0.0, 0.5)	C_{2h}
D	(0.0, 0.5, 0.5)	C_{2h}
D1	(0.0, 0.5, -0.5)	C_{2h}
E	(-0.5, 0.5, 0.5)	C_{2h}
H	(-0.48813803, 0.0, 0.51345699)	C_s
H1	(-0.51186197, 0.0, 0.48654301)	C_s
H2	(-0.48813803, 0.0, -0.48654301)	C_s
M	(-0.48813803, 0.5, 0.51345699)	C_1
M1	(-0.51186197, 0.5, 0.48654301)	C_1
M2	(-0.48813803, 0.5, -0.48654301)	C_1
X	(-0.5, 0.0, 0.0)	C_{2h}
Y	(0.0, 0.0, 0.5)	C_{2h}
Y1	(0.0, 0.0, -0.5)	C_{2h}
Z	(0.0, 0.5, 0.0)	C_{2h}
WO_3		
Γ	(0.0, 0.0, 0.0)	C_{2h}
A	(-0.5, 0.5, 0.0)	C_{2h}
C	(-0.5, 0.0, 0.5)	C_{2h}
D	(0.0, 0.5, 0.5)	C_{2h}
D1	(0.0, 0.5, -0.5)	C_{2h}
E	(-0.5, 0.5, 0.5)	C_{2h}
H	(-0.49453391, 0.0, 0.51125689)	C_s
H1	(-0.50546609, 0.0, 0.48874311)	C_s
H2	(-0.49453391, 0.0, -0.48874311)	C_s
M	(-0.49453391, 0.5, 0.51125689)	C_1
M1	(-0.50546609, 0.5, 0.48874311)	C_1
M2	(-0.49453391, 0.5, -0.48874311)	C_1
X	(-0.5, 0.0, 0.0)	C_{2h}
Y	(0.0, 0.0, 0.5)	C_{2h}
Y1	(0.0, 0.0, -0.5)	C_{2h}
Z	(0.0, 0.5, 0.0)	C_{2h}
SnO_2		
Γ	(0.0, 0.0, 0.0)	D_{4h}
Z	(0.0, 0.0, 0.5)	D_{4h}

A	(0.5, 0.5, 0.5)	D_{4h}
M	(0.5, 0.5, 0.0)	D_{4h}
R	(0.0, 0.5, 0.5)	D_{2h}
X	(0.0, 0.5, 0.0)	D_{2h}
$MgTa_2O_6$		
Γ	(0.0, 0.0, 0.0)	D_{4h}
Z	(0.0, 0.0, 0.5)	D_{4h}
A	(0.5, 0.5, 0.5)	D_{4h}
M	(0.5, 0.5, 0.0)	D_{4h}
R	(0.0, 0.5, 0.5)	D_{2h}
X	(0.0, 0.5, 0.0)	D_{2h}

The high-symmetry *k*-point which allows *p*-*d* coupling is highlighted in blue.

Table S3 Point groups of high-symmetry *k*-points in the first Brillouin zone of $Cs_{0.68}Ti_{1.83}O_4$

K point	Coordinate	Point group
$Cs_{0.68}Ti_{1.83}O_4$		
G	(0.0,0.0,0.0)	C_{2v}
R	(0.5,0.5,0.5)	C_{2v}
S	(0.5,0.5,0.0)	C_{2v}
T	(0.0,0.5,0.5)	C_{2v}
U	(0.5,0.0,0.5)	C_1
X	(0.5,0.0,0.0)	C_1
Y	(0.0, 0.5, 0.0)	C_{2v}
Z	(0.0, 0.0, 0.5)	C_{2v}

The high-symmetry *k*-point which allows *p*-*d* coupling is highlighted in blue.

Table S4 Mean (absolute) differences in the total energies and bandgaps calculated from the 117 B/N-codoping anatase TiO_2 configurations.

	Mean difference in total energy (meV/atom)	Mean absolute difference in total energy (meV/atom)	Mean difference in bandgap (eV)	Mean absolute difference in bandgap (eV)
B3a minus B4	-4.03	4.03	-0.03	0.04
B3b minus B4	1.26	1.69	-0.03	0.04

B4 represents doping configurations with four-coordinated B. B3a and B3b represent doping configurations with three-coordinated B in the basal plane and in the plane normal to the basal plane. The calculation method for the results obtained in this table is as follows: Given one of the 117 doping configurations, termed as B4, the corresponding most stable doping configurations with a three-coordinated B in the basal plane is termed as B3a. The energy/bandgap difference "B3a minus B4" is calculated by subtracting the energy/bandgap of the B4 from that of the B3a. The mean value of 117 such (absolute) energy/bandgap differences is the "mean (absolute) difference in energy/bandgap". Similarly, "B3b minus B4" were calculated.

Table S5 Performances of different ML models in predicting the bandgap values of B/N-codoping configurations in anatase TiO_2.

CGCNN model									
Model ID	# of model parameters	feature length	Init method[a]	max # of neighbors	# of conv layer	RMSE (eV)			MAE (eV)
						Train	Test	Identified 20	Identified 20
cfg-1	5041	16	Random	12	2	0.04	0.06	0.12	0.10
cfg-2	5041	16	Random	24	2	0.03	0.05	0.10	0.09
cfg-3	5041	16	Random	30	2	0.03	0.05	0.10	0.08
cfg-4	6257	16	Pretrained	24	2	0.03	0.06	0.16	0.14
cfg-5	7281	16	Random	24	3	0.02	0.06	0.18	0.13
cfg-6	14145	32	Random	24	2	0.02	0.05	0.10	0.08
Dexp2 structure descriptor + XGBoost model									
Model ID	# of model parameters	Max value of n[b]				RMSE (eV)			MAE (eV)
		x direction	y direction	z direction		Train	Test	Identified 20	Identified 20
cfg-1	5	1	1	1		0.02	0.07	0.19	0.17
cfg-2	5	2	2	2		0.03	0.08	0.22	0.20
cfg-3	5	3	3	3		0.03	0.07	0.22	0.21
This work									
Model ID	# of model parameters	No hyperparameters				RMSE (eV)			MAE (eV)
						Train	Test	Identified 20	Identified 20
cfg-1	5					**0.01**	**0.05**	**0.04**	**0.03**

[a]Init method refers to the method for generating the initial representations of elements in a CGCNN model. "Random" refers to generate a 4-dimensional random array for each element, "Pretrained" refers to using the representations read from the "atom_init.json" provided in the cgcnn package.

[b]Max value of n refers to the power in the Dexp2 descriptor.