Supplementary Information: Complex Ga_2O_3 Polymorphs Explored by Accurate and General-Purpose Machine-Learning Interatomic Potentials

Junlei Zhao,[1] Jesper Byggmästar,[2,3] Huan He,[2,4] Kai Nordlund,[2,5] Flyura Djurabekova,[2,5] and Mengyuan Hua[1]

[1]*Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen 518055, China*

[2]*Department of Physics, University of Helsinki, P.O. Box 43, FI-00014, Finland*

[3]*FCAI: Finnish Center for Artificial Intelligence, University of Helsinki, P.O. Box 43, FI-00014, Finland*

[4]*School of Nuclear Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China*

[5]*Helsinki Institute of Physics, University of Helsinki, P.O. Box 43, FI-00014, Finland*

CONTENTS

Appendix A: Convergence test of DFT calculations

The absolute convergence tests on the plane-wave energy cutoff "ENCUT", and the k-mesh grid density are shown in Fig. S1. A 20-atom monoclinic β-phase Ga_2O_3 unit cell (a = 12.460 Å, b = 3.086 Å, c = 5.879 Å, β = 103.68°) is used for the tests. It is shown that the plane-wave energy cutoff is a major source of uncertainty when set below 700 eV. The default energy cutoff in the VASP program will lead to an uncertainty in energy larger than 6 meV/atom. A 2×8×4 k-mesh grid should converge the energy difference well below 1 meV/atom. The parameters used for sampling the electronic system in this work should bring the total energy uncertainty down to less than 1 meV/atom.

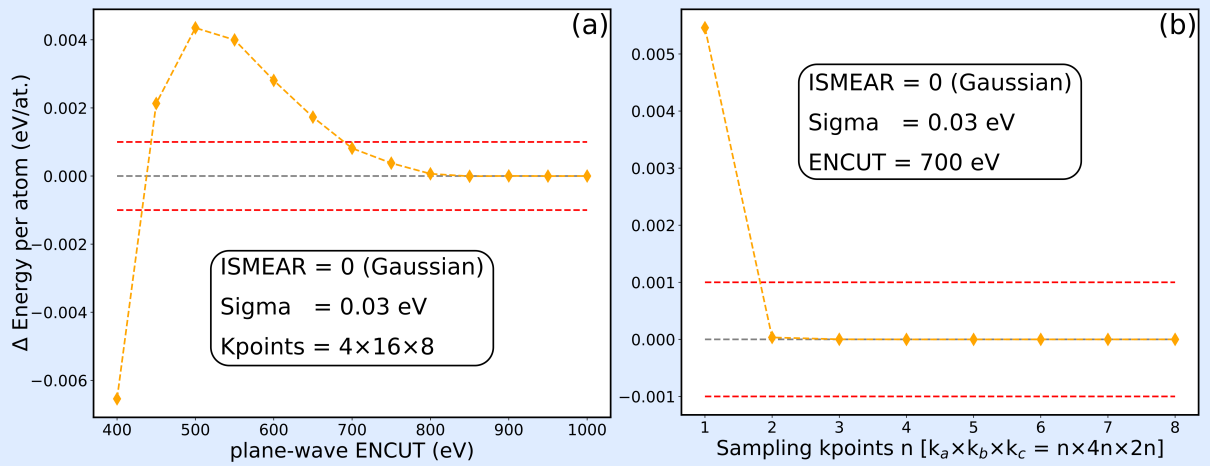

Figure S1. Total energy convergence tests on (a) plane-wave kinetic energy cutoff, "ENCUT", and (b) k-mesh grid with the Gaussian smearing, "ISMEAR=0" and the energy smearing width "SIGMA" of 0.03 eV, respectively. The zero-energy lines are aligned for (a) the ENCUT=1000 eV and 4×16×8 grid, and (b) ENCUT=700 eV and 8×32×16 grid, respectively, to emphasize the energy difference per atom. The red dashed lines are drawn at ± 10^{-3} eV/atom as a threshold of energy convergence.

Appendix B: Construction of the database

The generalized database is constructed with four main categories of structures:

(i). Static single-point DFT samplings of optimized cells, which are used to find the exact local minima of the potential energy surface.

(ii). *Ab initio* molecular dynamics (AIMD) at different temperatures and strain, which are used for fitting the correct thermal vibration and phonon dispersion of bulk systems at the near-equilibrium states.

(iii). Random active training with ML-predicted configurations recalculated by DFT.

(iv). Specifically selected configurations with high-energy repulsive interactions, such as close-3b phase, compressed O clusters and trimer O_3, to correct the initialle ML-predicted non-physical negative potential energy at the short-distance range.

Fig. S2 illustrates the scattered plot of the energy per atom versus the volume per atom for all the condensed fully-periodic DFT configurations. This plot can give a quick measure of the diversity of the DFT database. A more experimentally-relevant regime with energy ranging $-6.0 \sim -4.5$ eV/atom and volume ranging from $7 \sim 17$ Å^3 is zoomed.

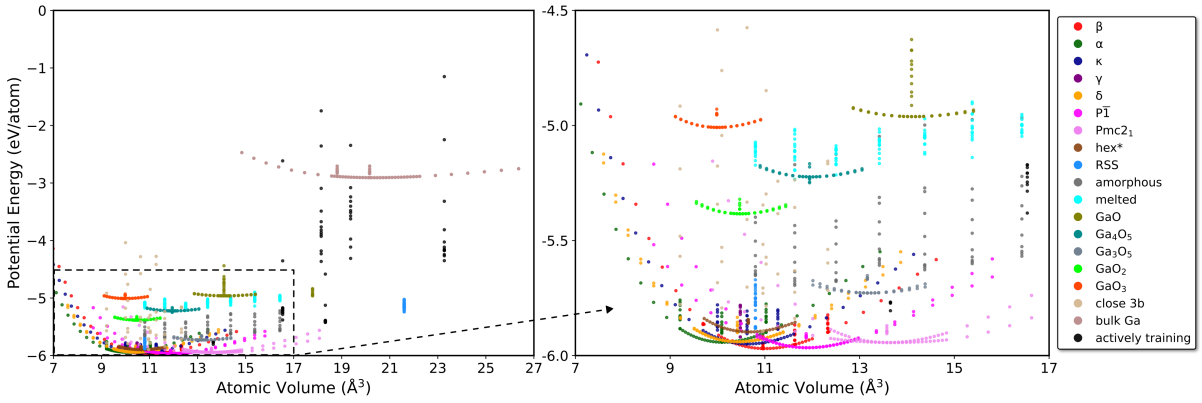

Figure S2. Energy versus atomic volume for all the fully-periodic configurations. The scattered colors indicate the different configuration types.

The detailed configuration types and their regularization (σ) are listed in Table S1. Note that the unit of the virial tensor is in (eV/Å^3)·(volume of cell in Å^3)=eV. As a well-tested generalized GAP training setup, the σ values of force and virial should be ten-times larger

than the one of the corresponding energy. For non-stoichiometric GaO_x configurations, a systematic test set of training with different σ was done. The current set results in the best balanced accuracy for both vital Ga_2O_3 polymorphs and GaO_x configurations.

Table S1. Configuration types and their specific regularization (σ)) corresponding to energy, force, virial tensor. The isolated Ga/O atoms are used as global references to calibrate the binding energies of all the other multi-atoms configurations.

Configuration types	Specific regularization (σ)		
	Energy (eV)	Force (eV/Å)	Virial (eV)
isolated Ga/O atoms	0.0001	0.010	0.010
β phase	0.0020	0.020	0.020
α phase	0.0020	0.020	0.020
κ phase	0.0020	0.020	0.020
γ phase	0.0020	0.020	0.020
δ phase	0.0020	0.020	0.020
$P\bar{1}$ phase	0.0020	0.020	0.020
$Pmc2_1$ phase	0.0020	0.020	0.020
hex* phase	0.0020	0.020	0.020
GaO	0.0035	0.035	0.035
Ga_4O_5	0.0035	0.035	0.035
Ga_3O_5	0.0035	0.035	0.035
GaO_2	0.0035	0.035	0.035
GaO_3	0.0035	0.035	0.035
pure Ga	0.0050	0.050	0.050
amorphous phase	0.0050	0.050	0.050
melted phase	0.0100	0.100	0.100
random structure search	0.0050	0.050	0.050
close-3b phase	0.0050	0.050	0.050
active training (O clusters)	0.0100	0.100	0.100
trimer O_3	0.0050	0.050	0.050
dimer Ga-Ga/Ga-O/O-O	0.0050	0.050	0.050

Appendix C: Training details of soapGAP and tabGAP

In both soapGAP and tabGAP training, we always use a twobody (2b) descriptor, $\mathbf{q}_{2\mathrm{b}}$, which is simply the distance between the center atom i and its neighboring atoms j:

$$\mathbf{q}_{i,2\mathrm{b}} = r_{ij} f_{\mathrm{cut}}(r_{ij}), \tag{S1}$$

where the finite cutoff function, $f_{\mathrm{cut}}(r_{ij})$, is, as conventional in classical empirical potentials, used for computational efficiency and linear scaling, and here amended to the 2b descriptor in the form:

$$f_{\mathrm{cut}}(r_{ij}) = \begin{cases} 1, & r_{ij} \leqslant r_{\mathrm{cut}} - \Delta r \\ \left[\cos\left(\frac{r_{ij} - r_{\mathrm{cut}} + \Delta r}{\Delta r}\right)\right]/2, & r_{\mathrm{cut}} - \Delta r < r_{ij} \leqslant r_{\mathrm{cut}} \\ 0, & r_{ij} > r_{\mathrm{cut}} \end{cases} \tag{S2}$$

where the cutoff range is between $r_{\mathrm{cut}} - \Delta r$ and r_{cut}. The cosine-function transition guarantees continuity of the function and its first-order derivative.

In the training of soapGAP, the manybody smooth overlap of atomic positions (SOAP) descriptor $\mathbf{q}_{\mathrm{SOAP}}$ (for detailed SOAP formula, refer to Refs. 1, 2) are used together with (2b) descriptor. In addition, for high-energy radiation damage simulation, we add an external pair potential, V_{pairs}, to better reproduce the Pauling repulsion at very short atomic distances (see Appendix C for details). Therefore, the total energy of N atoms in soapGAP is given by:

$$\begin{aligned} E_{\mathrm{tot.}} = & \sum_{i,j>i}^{N} V_{\mathrm{pairs}}(r_{ij}) \\ & + \delta_{2\mathrm{b}}^2 \sum_{i}^{N_{\mathrm{pairs}}} \sum_{s}^{M_{\mathrm{sparse}}^{2\mathrm{b}}} \alpha_{s,2\mathrm{b}} K_{2\mathrm{b}}(\mathbf{q}_{i,2\mathrm{b}}, \mathbf{q}_{s,2\mathrm{b}}) \\ & + \delta_{\mathrm{SOAP}}^2 \sum_{i}^{N} \sum_{s}^{M_{\mathrm{sparse}}^{\mathrm{SOAP}}} \alpha_{s,\mathrm{SOAP}} K_{\mathrm{SOAP}}(\mathbf{q}_{i,\mathrm{SOAP}}, \mathbf{q}_{s,\mathrm{SOAP}}). \end{aligned} \tag{S3}$$

The δ^2 factors are user-defined hyperparameters used as energy-scaling prefactors.

In the training of tabGAP, instead of the high-dimensional SOAP descriptor, we employ a threebody (3b) descriptor, $\mathbf{q}_{3\mathrm{b}}$ [3], together with a simple scalar density descriptor based

on the embedded atom method (EAM) [4, 5]. The 3b descriptor is a three-component vector with all information about the bond lengths and angle:

$$\mathbf{q}_{i,3\mathrm{b}} = [r_{ik} + r_{ij}, (r_{ik} - r_{ij})^2, r_{jk}] \cdot f_{\mathrm{cut}}(r_{ij}) f_{\mathrm{cut}}(r_{ik}) \tag{S4}$$

where the cutoff function is the same as Eq. S2. The generalized many-element form of the EAM descriptor is:

$$\mathbf{q}_{i,\mathrm{EAM}} = \sum_{j \neq i} \kappa_{ij} \phi(r_{ij}), \tag{S5}$$

where κ_{ij} is a prefactor that depends on the chemical species of i and j, and $\phi(r_{ij})$ is a scalar function for the pair-density contribution. More detailed formulation of the EAM descriptor can be found in Ref. 5. Therefore, similar to Eq. S3, the total energy of N atoms in this "low-dimensional" GAP is then given by:

$$\begin{aligned} E_{\mathrm{tot.}} = & \sum_{i,j>i}^{N} V_{\mathrm{pairs}}(r_{ij}) \\ & + \delta_{2\mathrm{b}}^2 \sum_{i}^{N_{\mathrm{pairs}}} \sum_{s}^{M_{\mathrm{sparse}}^{2\mathrm{b}}} \alpha_{s,2\mathrm{b}} K_{2\mathrm{b}}(\mathbf{q}_{i,2\mathrm{b}}, \mathbf{q}_{s,2\mathrm{b}}) \\ & + \delta_{3\mathrm{b}}^2 \sum_{i}^{N_{\mathrm{triplets}}} \sum_{s}^{M_{\mathrm{sparse}}^{3\mathrm{b}}} \alpha_{s,3\mathrm{b}} K_{3\mathrm{b}}(\mathbf{q}_{i,3\mathrm{b}}, \mathbf{q}_{s,3\mathrm{b}}) \\ & + \delta_{\mathrm{EAM}}^2 \sum_{i}^{N} \sum_{s}^{M_{\mathrm{sparse}}^{\mathrm{EAM}}} \alpha_{s,\mathrm{EAM}} K_{\mathrm{EAM}}(\mathbf{q}_{i,\mathrm{EAM}}, \mathbf{q}_{s,\mathrm{EAM}}) \end{aligned} \tag{S6}$$

The detailed hyperparameters used for constructing the descriptors, the kernel functions, and training the soapGAP and low-dimensional GAP are summarized in Table S2, and S3, respectively. To make the final tabulated version of the low-dimensional GAP (the tabGAP), the 2b- and 3b-energy terms in Eq. S6 are further mapped onto 1D and 3D grids. The tabGAP energy is then computed efficiently using cubic splines, yielding a two-orders-of-magnitude speedup compared to the original low-dimensional GAP. The EAM-energy term follows a conventional format as a tabulated Finnis-Sinclair EAM potential [6] file with 1D splines for the pair density function and embedding energies. The code for generating tabGAP potential files and using them in LAMMPS is available in Ref. 7.

For more details about the construction of the Gaussian approximation potential, we

Table S2. Hyper-parameters used for training the 2b+SOAP soapGAP.

Universal prefactors and cutoffs	
δ_{2b} & δ_{SOAP}	2.0
r_{cut}^{2b} & r_{cut}^{SOAP} (Å)	5.0
Δr^{2b} & Δr^{SOAP} (Å)	1.0
SOAP descriptor	
ζ	4
$\sigma_{at.}$ (Å)	0.5
n_{max}	8
l_{max}	8
SOAP M_{sparse}^{SOAP}	8000
SOAP sparse method	"cur points"
SOAP kernel function	"dot product"
2b descriptor	
2b M_{sparse}^{2b}	20
2b $\theta_{uniform}$	1.0
2b sparse method	"uniform"
2b kernel function	"ard se"

refer the reader to Refs. 1, 2, 8. The training processes are performed using the QUantum mechanics and Interatomic Potentials (QUIP) package [9]. The testing validation and the molecular dynamics simulations are done using Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) package [10]. Two exemplary inputs for using soapGAP and tabGAP in LAMMPS are shown below:

- To use soapGAP ("xml_label" should be updated accordingly):

```
pair_style  quip
pair_coeff  * *  gap_GaO_20220921.xml &
"Potential xml_label=GAP_2022_9_22_480_23_27_2_249" 31 8
```

- To use tabGAP:

```
pair_style hybrid/overlay eam/fs tabgap
pair_coeff * * eam/fs Ga-O.eam.fs Ga O
pair_coeff * * tabgap Ga-O.tabgap Ga O no yes
```

Table S3. Hyper-parameters used for training the low-dimensional 2b+3b+EAM GAP. Note that the Finnis-Sinclair-like pair density function [5, 6] does not require a cutoff function, thus no Δr^{EAM} is defined.

Universal prefactors and cutoffs	
δ_{2b}	2.0
δ_{3b} & δ_{EAM}	1.0
$r_{\mathrm{cut}}^{\mathrm{2b}}$ & $r_{\mathrm{cut}}^{\mathrm{EAM}}$ (Å)	5.0
Δr^{2b} (Å)	1.0
$r_{\mathrm{cut}}^{\mathrm{3b}}$ (Å)	4.0
Δr^{3b} (Å)	0.7
2b descriptor	
2b $M_{\mathrm{sparse}}^{\mathrm{2b}}$	20
2b $\theta_{\mathrm{uniform}}$	1.0
2b sparse method	"uniform"
2b kernel function	"ard se"
3b descriptor	
3b $M_{\mathrm{sparse}}^{\mathrm{3b}}$	300
3b $\theta_{\mathrm{uniform}}$	1.0
3b sparse method	"uniform"
3b kernel function	"ard se"
EAM descriptor	
EAM $M_{\mathrm{sparse}}^{\mathrm{EAM}}$	20
EAM $\theta_{\mathrm{uniform}}$	1.0
EAM sparse method	"uniform"
EAM kernel function	"ard se"
EAM pair function	"FSgen"
EAM pair function order	3
EAM prefactor "mode"	"FSgen"

Appendix D: High-energy repulsive potential

As mentioned in Appendix B, the interatomic repulsion at extremely short distance is considered by the external pair potential in the form of a refitted Ziegler-Biersack-Littmarck (ZBL) potential [11] which has the following original form:

$$V_{\text{pairs}}(r_{ij}) = \frac{1}{4\pi\varepsilon_0}\frac{Z_i Z_j e^2}{r_{ij}}\phi(r_{ij}/a) f_{\text{cut}}^{\text{pairs}}(r_{ij}), \tag{S7}$$

where a is in the semi-empirical form:

$$a = \frac{0.46848}{Z_i^{0.23} + Z_j^{0.23}}, \tag{S8}$$

where the cutoff function, $f_{\text{cut}}^{\text{pairs}}(r_{ij})$, is in the form:

$$f_{\text{cut}}^{\text{pairs}}(r_{ij}) = \begin{cases} 1, & r_{ij} \leqslant r_1 \\ 1 - \chi^3(6\chi^2 - 15\chi + 10), & r_1 \leqslant r_{ij} \leqslant r_2 \\ 0, & r_{ij} \geqslant r_2 \end{cases} \tag{S9}$$

where $\chi = (r_{ij} - r_1)/(r_2 - r_1)$. The different V_{pairs} cutoff ranges are chosen depending on the pair types. Moreover, we refit the screening functions, $\phi(r_{ij}/a)$, for each pair using the all-electron DMol-DFT data calculated from Ref. 12 with a double-exponential function in the form:

$$\phi(x) = A_1 \exp(-B_1 x) + (1 - A_1)\exp(-B_2 x) \tag{S10}$$

where A_i and B_i are the fractional energies and the damping factors of the i-th component. The fitted parameters are summarized in Table S4.

Table S4. Fitted parameters of the screening function in the form of Eqs. S10 and S9

Parameters	Ga-Ga	Ga-O	O-O
A_1	0.58004	0.44723	0.27664
B_1	1.22784	1.65924	2.04820
B_2	0.43165	0.50402	0.59505
r_1	1.5	0.75	0.50
r_2	2.5	1.5	1.0

The refitted pair potentials, V_{pairs}, at extremely short distances are shown in Fig. S3. The standard empirical ZBL potentials and energies calculated using plane-wave (PW) DFT are plotted for reference. At the energy range from 1 MeV to 1 keV, the pair potentials agree with the DMol-DFT data and the ZBL potentials very well, while the ZBL potentials deviate from the *ab initio* data (see the inset figures in Fig. S3), at the mid-range from 1 keV to 10 eV. The cutoff functions lead to smooth transition region in between the high-energy and near-equilibrium atomic distances.

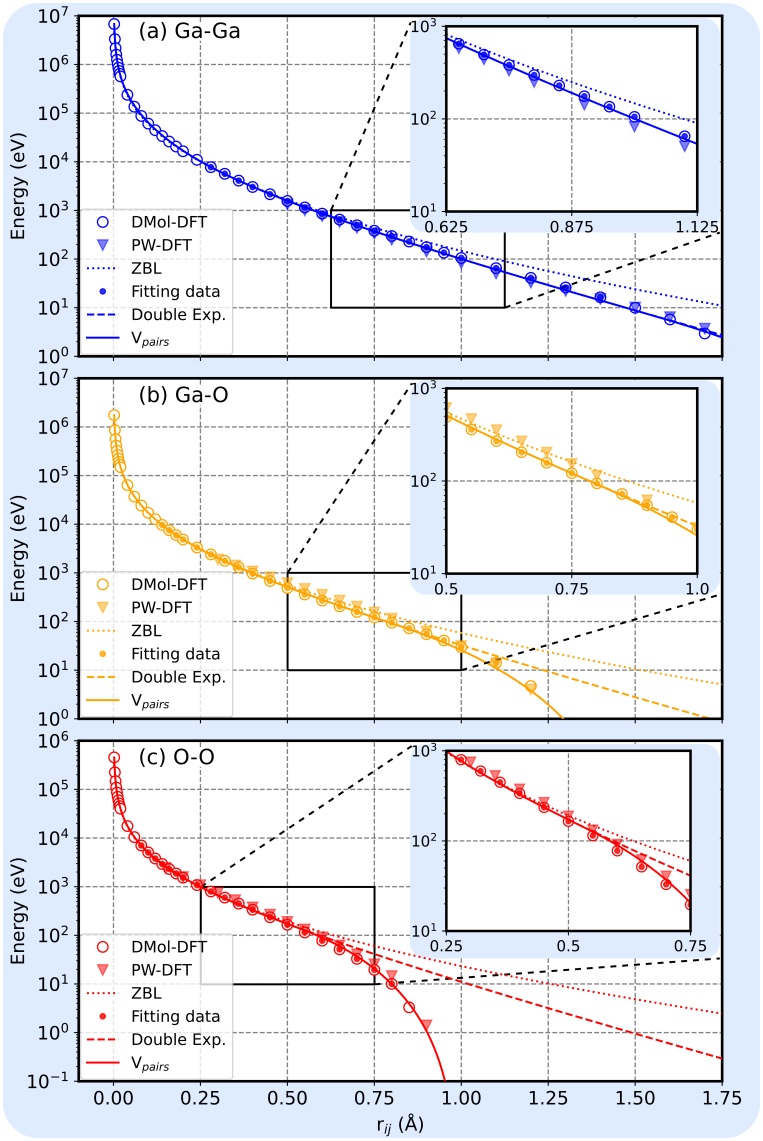

Figure S3. The log-linear plot of the repulsive pair potentials at high-energy range. (a) Ga-Ga; (b) Ga-O; (c) O-O. The inset figures are zoomed in the energy range from 1 keV to 10 eV. The large dots indicate the DMol-DFT data points used for fitting the Eq. S10.

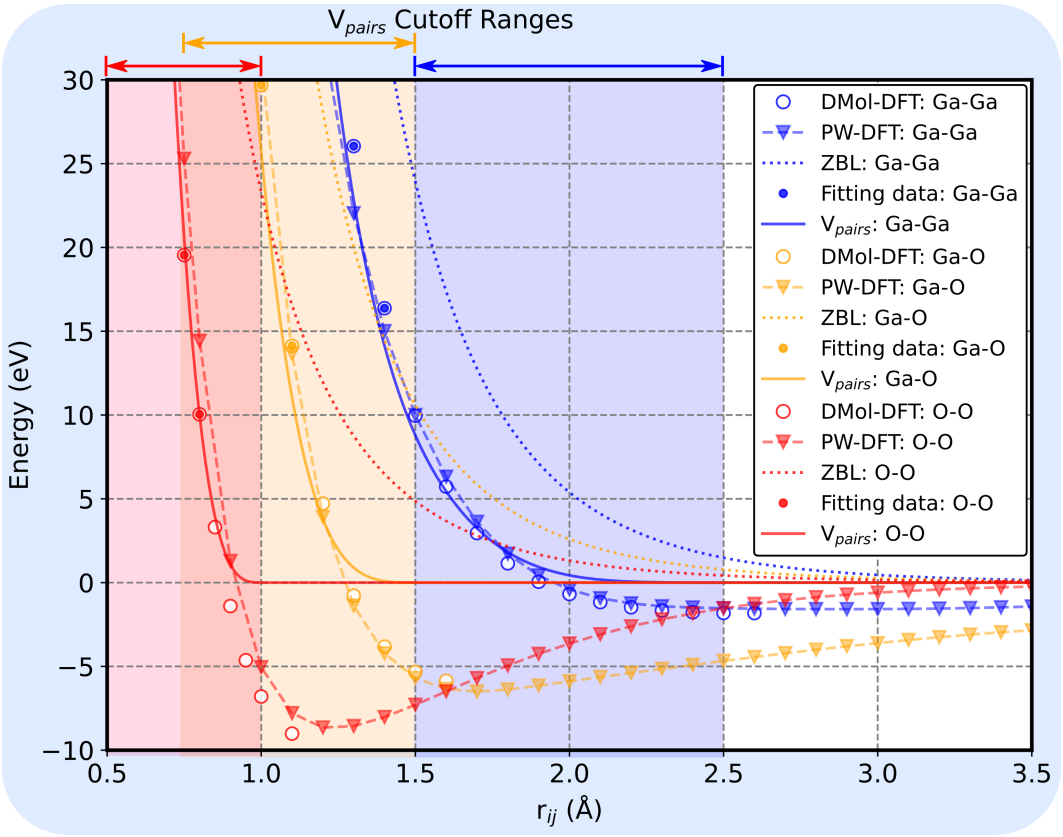

Figure S4. The linear-linear plot of the repulsive pair potentials at near-equilibrium regions. The color shadowed regions illustrate the corresponding ranges of the cutoff functions.

Appendix E: Chemical potential of non-stoichiometric systems

As shown in Fig. S5, we test the stabilities of the five GaO_x configurations against the reference energies of physical mixing, pure Ga, global minimum β-Ga_2O_3 and pure gas-phase O_2/O_3. The soapGAP predicts smaller energy errors for all five calculated phases compared to the tabGAP. Both soapGAP and tabGAP were tested with Metropolis Monte Carlo atom-swapping runs and no non-physical segregation was seen starting from all the five Ga_2O_3 polymorphs, in neither bulk nor open-surface systems. We note that the energies of the pure O_2 and O_3 is reversed in the current version of tabGAP, which is a consequence of using only low-dimensional descriptors. The low dimensionality sets limits on the flexibility, similar to parametric analytical potentials, so that in our case oxygen molecules cannot be described perfectly without significantly sacrificing the accuracy of the important bulk structures. For our tabGAP, we have balanced the 2b/3b/EAM energy terms to accurately fit of the most important Ga_2O_3 bulk polymorphs.

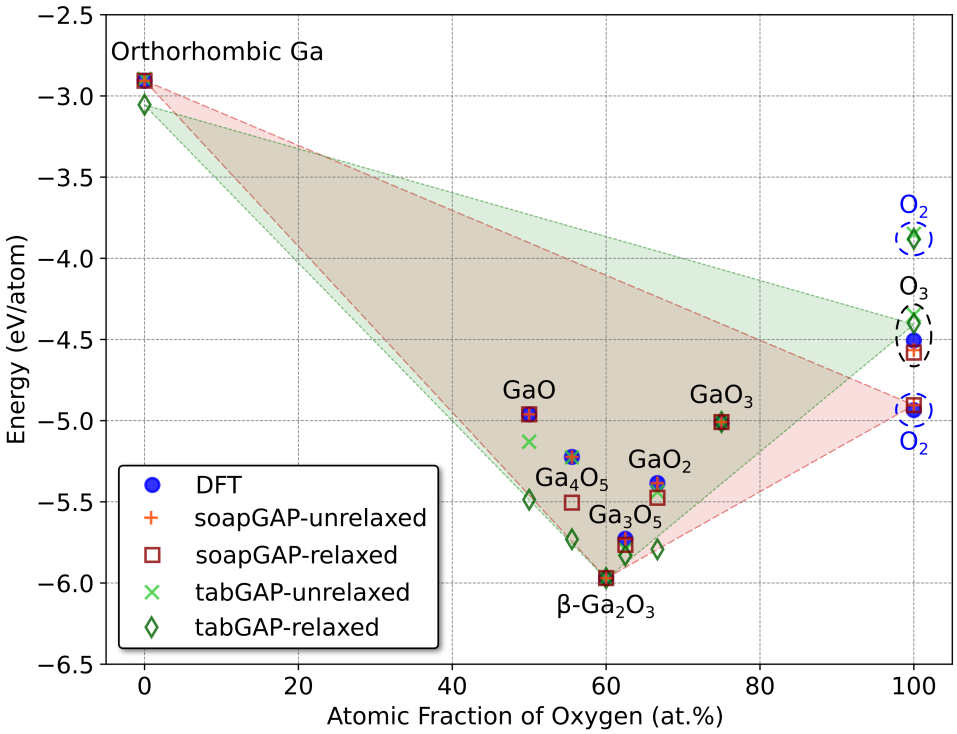

Figure S5. Chemical potentials of non-stoichiometric phases against the references of pure Ga/O, and β-Ga_2O_3 phases. The red and green triangles indicate the reference energies of physical mixing for soapGAP and tabGAP, respectively. The unrelaxed structures are the same as the DFT calculated ones, while the relaxed structures are the soap/tabGAP-predicted local minima relaxed with full degree of freedom in both internal atom positions and external lattice stresses.

Appendix F: Elastic properties and melting point

The GAP-predicted bulk moduli (B_0), pressure derivatives of bulk moduli ($B_0^{'}$), of the four polymorphs are listed in Table S5. The γ-phase is not examined here due to its highly complex nonequivalent configurations of defective spinel structures. In general, the soapGAP-predicted bulk moduli are in good agreement with the DFT data, while the tabGAP-predicted B_0 of the β-phase is slightly overestimated,, which leads to a higher value than that of the κ-phase.

Table S5. Bulk moduli B_0 (in GPa) and their pressure derivatives $B_0^{'}$ of the $\beta/\kappa/\alpha/\delta$ polymorphs predicted by GGA-DFT, soapGAP and tabGAP. The trend is in agreement with the *ab initio* data calculated with GGA-AM05 *xc* functional [13], where the order of B_0 is $\beta < \kappa < \delta < \alpha$.

	GGA-DFT	soapGAP	tabGAP
β-phase (Monoclinic, $C2/m$)			
B_0	156.0	157.3	177.5
$B_0^{'}$	2.600	2.961	2.813
κ-phase (Orthorhombic, $Pna2_1$)			
B_0	167.9	171.7	167.6
$B_0^{'}$	4.347	4.325	3.262
α-phase (Trigonal, $R\bar{3}c$)			
B_0	197.5	197.5	193.7
$B_0^{'}$	5.268	2.918	6.089
δ-phase (Cubic, $Ia\bar{3}$)			
B_0	177.0	181.2	183.4
$B_0^{'}$	6.777	3.945	7.123

The tabGAP-predicted melting point of the β-phase is $\sim 2020 \pm 10$ K as shown in Fig. S6 from the heating process. The NPT MD stimulation starts from an initial bulk 1280-atom β-phase cell with a block of 20 atoms removed as the heterogeneous nucleation seed of melting. In the heating process, the temperature is increased 10 K every 200 ps from 2000 K. A fast solid-liquid transition occurs around 2010 $\sim$ 2030 K in three independent runs (here the 2010-K case is shown in Fig. S6). In the cooling process, the temperature is reduced 50 K every 100 ps to 600 K finally. As the pressure control (kept at 0 bar) is uncoupled in the three dimensions, the shape of the liquid cell changes freely (from 500 ps to 2000 ps in Fig. S6) until the mobility of the atoms is low (solidified) at the low temperature.

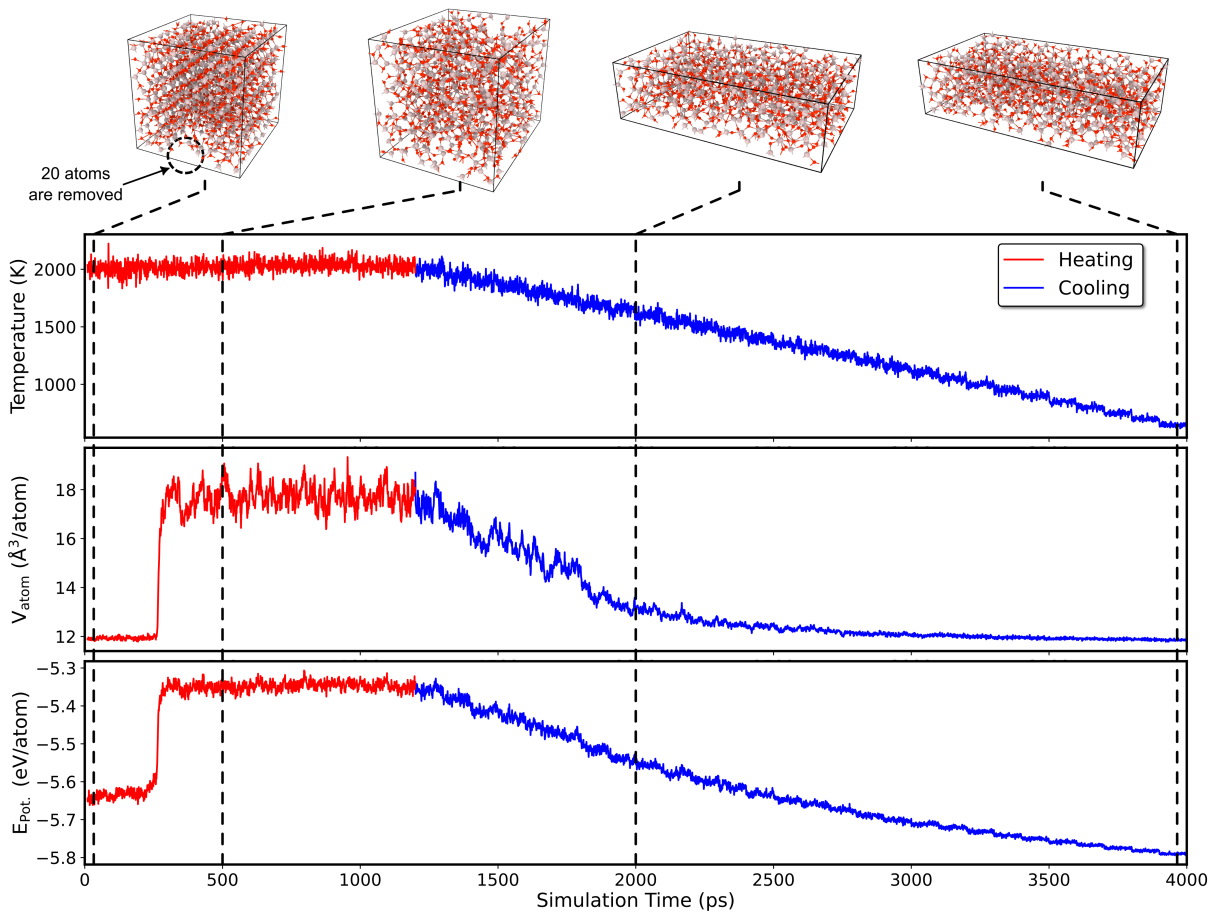

Figure S6. The tabGAP MD simulation of the melting and quenching process. The evolution of the atomic configurations, temperature, atomic volume (V_{atom}), and potential energy ($E_{\mathrm{pot.}}$) are shown.

Appendix G: Computational performance

We test the computational speed of the soapGAP and tabGAP on a supercomputer node consisting of 2 Xeon-Gold-6148 CPUs (clocked at 2.4 GHz, 20 cores per CPU). The testing simulations are run with the same 320-atom β-phase Ga_2O_3 system, since the same cell can be effectively run with AIMD on a single 40-core node as a direct reference. The classical MD is run with LAMMPS-23Jun2022 version built with Intel-MPI (v2018.4), and the AIMD is run with VASP-6.3.1 version built with the same MPI module.

The average soapGAP-MD loop time is $\sim 10^4$ s for 25000 steps, and tabGAP-MD ~ 25 s for 25000 steps. Note that the fraction of inter-processor communication is about $20 \sim 25$ % of the total running time owing to the small system. The AIMD (single-Γ sampling, ENCUT = 550 eV, energy threshold 10^{-5} eV, blocked-Davidson algorithm) running time is ~ 20000 s for 100 AIMD steps.

Therefore, in this single-point estimation, the soapGAP and tabGAP are boosted by 500 and 2×10^5 times in computing speed compared to the *ab initio* method, respectively. We note that DFT scales with the system size as $\sim \mathcal{O}(N^3)$ and MD as $\sim \mathcal{O}(N)$ with much lower computing memory consumption, so in a large-scale system of $10^3 \sim 10^5$ atoms, the GAP-MD runs are exceedingly more efficient than the AIMD runs.

[1] A. P. Bartók, R. Kondor, and G. Csányi, Phys. Rev. B **87**, 184115 (2013).

[2] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, Chem. Rev. **121**, 10073 (2021).

[3] A. P. Bartók and G. Csányi, Int. J. Quantum Chem. **115**, 1051 (2015).

[4] S. M. Foiles, M. I. Baskes, and M. S. Daw, Phys. Rev. B **33**, 7983 (1986).

[5] J. Byggmästar, K. Nordlund, and F. Djurabekova, Phys. Rev. Mater. **6**, 083801 (2022).

[6] M. W. Finnis and J. E. Sinclair, Philos. Mag. A **50**, 45 (1984).

[7] https://gitlab.com/jezper/tabgap.

[8] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).

[9] Quip–quantum mechanics and interatomic potentials, https://github.com/libatoms/quip.

[10] S. Plimpton, J. Comp. Phys. **117**, 1 (1995).

[11] J. F. Ziegler, J. P. Biersack, and U. Littmark, *The Stopping and Range of Ions in Matter* (Pergamon, New York, NY, 1985).

[12] K. Nordlund, N. Runeberg, and D. Sundholm, Nucl. Instrum. Methods Phys. Res., Sect. B **132**, 45 (1997).

[13] J. Furthmüller and F. Bechstedt, Phys. Rev. B **93**, 115204 (2016).