Electronic Supplementary Information to the article "Ordering and phase separation in Gd-doped ceria: A combined DFT, cluster expansion and Monte Carlo study"

Pjotrs A. Žguns,*[ab] Andrei V. Ruban[bc] and Natalia V. Skorodumova[ab]

Notice: The order of covered topics may differ from the order in the main article.

Regarding abbreviations: Here we use abbreviations for description of a set of structures:

S — supercell size,
V — number of Va,
R / N — number of relaxed / non-relaxed structures in the set,
a — lattice constant (in Å),
PS / PE — PBEsol / PBE exchange-correlation functional.

Label like S4_V16_R168_a_5.40_PS means the following set of structures: S4 – $4 \times 4 \times 4$ supercells, V16 – 16 Va (16 neutral oxygen vacancies in the supercell, thus also 32 Gd^{3+}), R168 – 168 structures (with ionic relaxations), a_5.40 – 5.402 Å large lattice constant (a_5.47 implies 5.470 Å), PS – PBEsol exchange-correlation functional.

[a] *Department of Physics and Astronomy, Uppsala University, Box 516, 75121 Uppsala, Sweden. E-mail: pjotr.zgun@gmail.com*
[b] *Department of Materials Science and Engineering, KTH Royal Institute of Technology, 10044 Stockholm, Sweden*
[c] *Materials Center Leoben Forschung GmbH, A-8700 Leoben, Austria*

1 Supplementary information to the section *Cluster Expansion: methodology, results and discussions*

1.a Choice of cluster interactions (CIs)

Here we start with the choice of cluster interactions (CIs) to be included in the Hamiltonian. For this purpose we use the largest set of structures (set_03: 210 structures, $3 \times 3 \times 3$ supercells, $x_{\mathrm{Gd}} = 0.1296$, $a = 5.402$ Å, PBEsol functional [see Appendix 1]). When reading this section, please, refer to Sections 3.2 and 3.3 in the main part of this paper.

The simplest choice: $\mathcal{H}_0$

The nomenclature provided in Section 3.2 (Tables 1–4) lists all possible structurally different and "mediated pair" CIs, arising from the complex structure of GDC (and some triangular CIs). However, it may not be necessary to consider all these complex CIs, hence we start with the simplest choice of CIs ($\mathcal{H}_0$), which approximates configurational energy as follows:

1) All structurally different pair CIs (with the same vector) and mediated pair CIs (with the same vector) are treated as one CI. (CIs corresponding to the cases a, b, c, d are considered as one CI, *e.g.* $J_{3a}^{\mathrm{VaVa}} \equiv J_{3b}^{\mathrm{VaVa}} \equiv J_{3c}^{\mathrm{VaVa}}$.)

2) $\mathcal{H}_0$ does not include three site CIs (see Table 4).

3) The largest possible cut-off radius is used. Since $3 \times 3 \times 3$ supercells are considered, the cut-off radius is $1.871a$ and GdVa CIs 9–11 are excluded (see Appendix 2 and Section 3.3).

Ways to systematically improve $\mathcal{H}_0$

In order to systematically improve $\mathcal{H}_0$, one can include:

a) S_i — structural differentiation for i-th CIs, *i.e.* between structurally different pair CIs.

b) C_i — complex differentiation between structurally different and mediated pair CIs, thus among a, b, c and d cases of i-th CIs (*i.e.* C_i is more advanced than S_i and includes it).

c) T — consider triangular three site CIs.

For example, let us consider $\langle \frac{1}{2}, \frac{1}{2}, \frac{1}{2} \rangle$ Va−Va CIs ($3a$: Va−empty−Va; $3b$: Va−Ce−Va, $3c$: Va−Gd−Va). S_3 implies a differentiation between $3a$ and $3b$, but $3c$ is not considered (counted as $3b$), while C_3 implies a differentiation among all $3a$, $3b$ and $3c$ cases.

Results

First, let us compare Hamiltonians $\mathcal{H}_0$ with $\mathcal{H}_1$ ($\mathcal{H}_0$ with Va−Va S_3) and $\mathcal{H}_2$ ($\mathcal{H}_0$ with Va−Va C_3). We number Hamiltonians with an unique subscript index. Table S1 shows that the Va−Va differentiation C_3 (among cases $3a$, $3b$, $3c$) is important: cross-validation (CV)

score[1] is systematically improved (becomes smaller) from $\mathcal{H}_0$ to $\mathcal{H}_1$, and further to $\mathcal{H}_2$. Last but not least, C_3 is necessary because J's are rather different: $J_{3a}^{\mathrm{VaVa}} = 0.26$ eV, $J_{3b}^{\mathrm{VaVa}} = 0.38$ eV, $J_{3c}^{\mathrm{VaVa}} = 0.57$ eV (these values are obtained using set_03 and $\mathcal{H}_5$, see below).

Table S1: Comparison of choices of CIs. CIs "added" to the simplest $\mathcal{H}_0$ are listed (see Tables 1–3 in the main paper). CV scores are based on the set_03.

Choice	Additional CIs			CV (meV/defect)
	Gd−Gd	Va−Va	Gd−Va	
$\mathcal{H}_0$	–	–	–	6.40
$\mathcal{H}_1$	–	$3b$	–	5.55
$\mathcal{H}_2$	–	$3bc$	–	4.78

Table S2: Further comparison of choices of CIs: trying to improve $\mathcal{H}_2$. CIs "added" to the simplest $\mathcal{H}_0$ are listed (see Tables 1–3 in the main paper). CV scores are based on the set_03.

Choice	Additional CIs			CV (meV/defect)
	Gd−Gd	Va−Va	Gd−Va	
$\mathcal{H}_3$	$4b$	$3bc$	–	4.79
$\mathcal{H}_4$	–	$3bc$	$4b$	4.75
$\mathcal{H}_5$	–	$3bc, 4b$	–	4.20
$\mathcal{H}_6$	–	$3bc, 4b, 7b$	–	4.20
$\mathcal{H}_7$	–	$3bc, 4b, 7b, 9bc$	–	4.17
$\mathcal{H}_8$	$4b, 6bc$	$3bc, 4b, 7b, 9bc, 11bc, 12bcd$	$4b$	4.28

The further influence of the complex differentiations is shown in Table S2. Additional consideration of $4b$ Gd−Gd or $4b$ Gd−Va does not affect the CV score, because these CIs are more distant and are rather weak compared to Va−Va $3a, 3b, 3c$. However, the consideration of $4b$ Va−Va decreases the CV score. Further consideration of mediated Va−Va CIs ($7b$ and $9b$, $9c$) and other structurally different (or mediated) CIs does not change the CV score significantly.

The influence of three site CIs was checked by including the smallest triangular clusters of four different kinds (3Gd, 2Gd&1Va, 1Gd&2Va, 3Va). The addition of these CIs to Hamiltonians, in fact, increases the CV score (compare Table S3 with Tables S1 & S2).

[1]The CV score is defined as

$$\mathrm{CV} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(E_i - E_i^{'}\right)^2}, \tag{1}$$

where index i runs over the n structures, E_i is the calculated energy and $E_i^{'}$ is the cluster expansion (CE) predicted energy based on CIs obtained using $n-1$ structures (all except i-th) as an input to CE (see A. van de Walle, *JOM*, 2013, **65**, 1523-1532). The E_i and $E_i^{'}$ are usually normalised per atom, but here we use the total number of Gd and Va in the supercell as the normalisation constant, because in our Hamiltonian only Gd and Va are interacting. Thus we use [CV score] = [meV/defect].

Table S3: Further comparison of choices of CIs: consideration of three site CIs. CIs "added" to the simplest $\mathcal{H}_0$ are listed (see Tables 1–4 in the main paper; T denotes three site CIs). CV scores are based on the set_03.

Choice	Additional CIs				CV (meV/defect)
	Gd−Gd	Va−Va	Gd−Va	T	
$\mathcal{H}_9$	–	$3bc$	–	1–4	4.89
$\mathcal{H}_{10}$	–	$3bc, 4b$	–	1–4	4.30
$\mathcal{H}_{11}$	$4b, 6bc$	$3bc, 4b, 7b, 9bc, 11bc, 12bcd$	$4b$	1–4	4.38

Therefore we do not account for triangular three site CIs and further use the choice $\mathcal{H}_5$ for CE.

1.b Effect of cut-off radius on cross validation (CV) score

Previous tests (Section 1.a) were performed with the fixed cut-off radius. Here we examine how CV score behaves upon the increase of the cluster expansion's (CE's) cut-off radius ($R_{\mathrm{CUT}}^{\mathrm{CE}}$). Fig. S1 shows that for both the $3\times3\times3$ and $4\times4\times4$ supercells (set_03 and set_11, respectively [see Appendix 1]) the CV score decreases upon the increase of $R_{\mathrm{CUT}}^{\mathrm{CE}}$.

Thus, in order to make an accurate CE, one has to include as much as possible long-range CIs, provided that the number of structures is sufficiently larger than the number of unknown CIs, enhancing the robustness of CE and guaranteeing the reduction of CV score upon increase of $R_{\mathrm{CUT}}^{\mathrm{CE}}$.

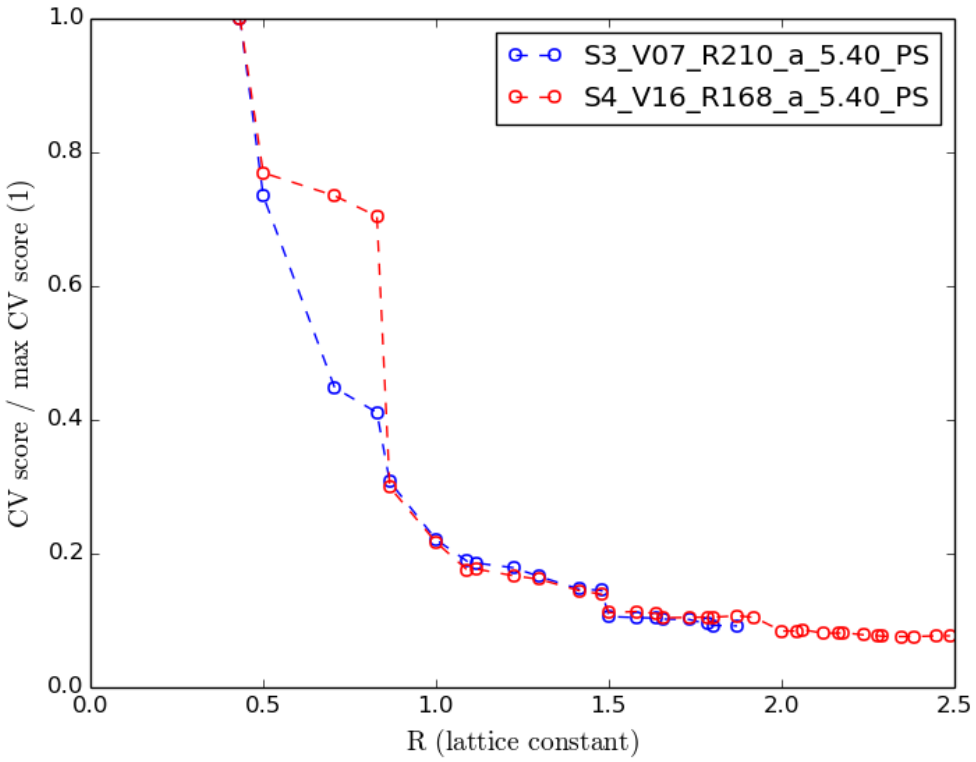

Figure S1: Normalised CV scores for $3\times3\times3$ and $4\times4\times4$ supercells. The CV scores for the most long-range cut-off radii are 4.2 meV/defect (0.3 meV/lattice site) and 2.4 meV/defect (0.2 meV/lattice site) for $3\times3\times3$ and $4\times4\times4$ cases, respectively.

1.c Weighting of structures: effect on CV score and CIs

Here we investigate the effect of structure weighting on the CV score and CIs. The goal is to increase a weight of low-energy structures, which are closer to the ground state. This might reduce the CV score and provide CIs more appropriate for description of the ground state in Monte Carlo.

The procedure is the following: we assign the statistical weight to the i-th structure $\propto \exp[-(E_i - E_0)/(k_\mathrm{B}T_\mathrm{w})]$, where E_i is the energy of the i-th structure, E_0 is the minimal energy in the set (energies are normalised by $N_\mathrm{defects} = N_\mathrm{Gd} + N_\mathrm{Va}$) and parameter T_w is a weighting temperature.

Fig. S2 shows that upon decrease of the weighting temperature, the CV score slightly decreases for low energy structures, while for high energy structures the CV score increases. Finally, weighting virtually does not affect the CIs (see Figs. S3 and S4) therefore we do not apply it.

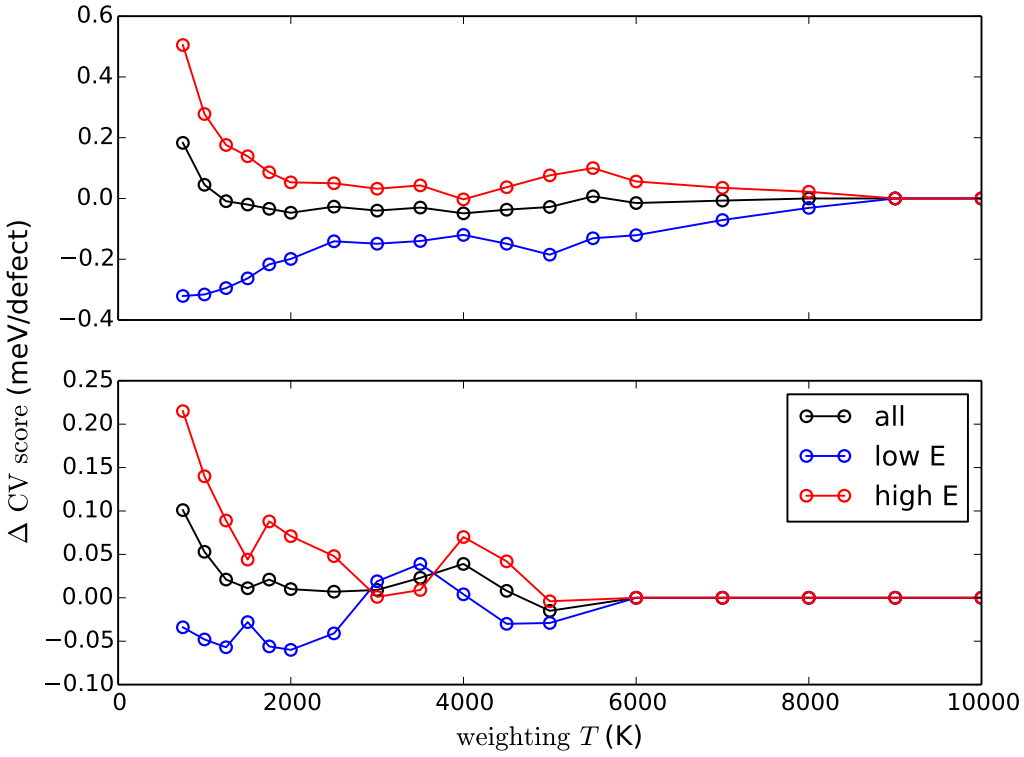

Figure S2: Change of CV score upon decrease of weighting temperature for: 1) all structures, 2) one half of all structures (with lowest energies), 3) the other half of all structures (with highest energies). Top panel: 210 structures, $3 \times 3 \times 3$ supercell, 7 Va. Bottom panel: 168 structures, $4 \times 4 \times 4$ supercell, 16 Va. In both panels: 5.402 Å lattice constant, PBEsol exchange-correlation functional.

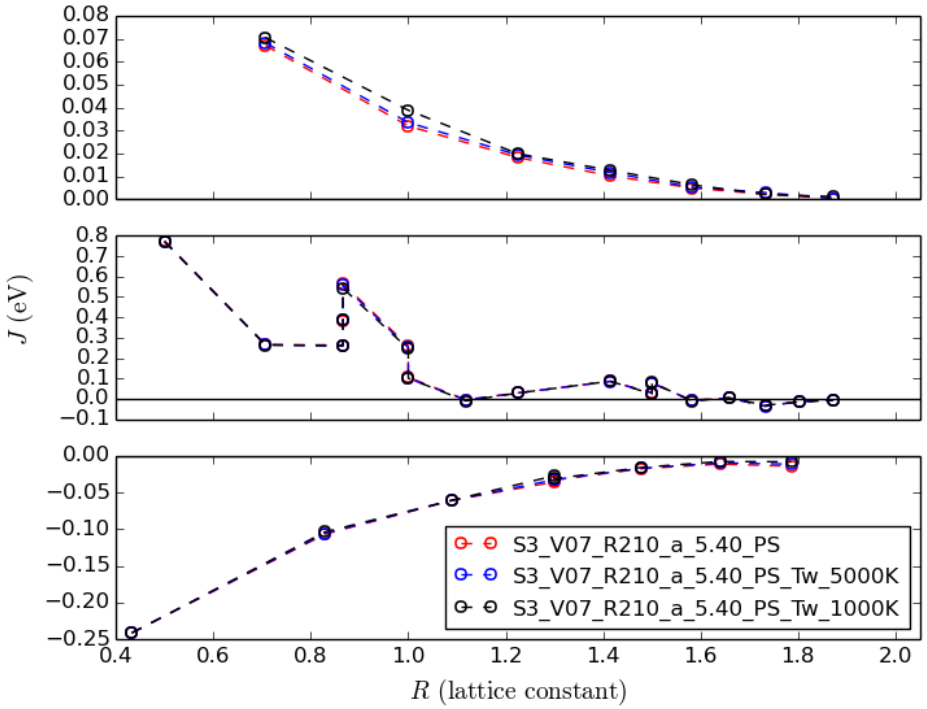

Figure S3: Effect of weighting on CIs. Tw specifies the weighting temperature in K. $3\times3\times3$ supercells. CIs (J's) *vs.* distance, from top to bottom: Gd−Gd, Va−Va, Gd−Va.

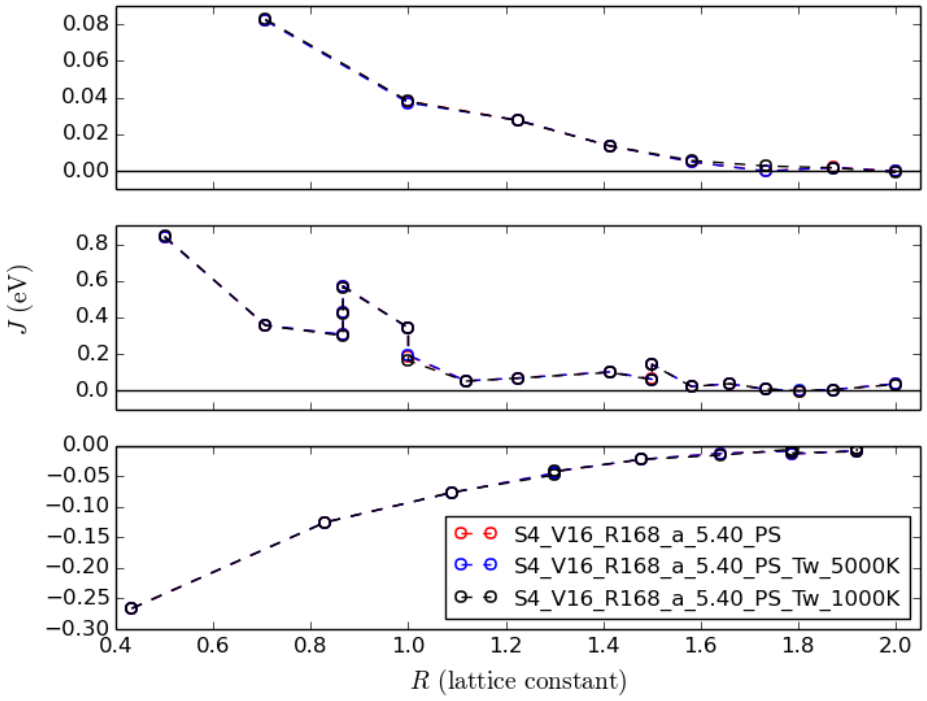

Figure S4: Effect of weighting on CIs. Tw specifies the weighting temperature in K. $4\times4\times4$ supercells. CIs (J's) *vs.* distance, from top to bottom: Gd−Gd, Va−Va, Gd−Va.

1.d Convergence of CIs with respect to the set size

Here we analyse the set of 210 structures ($3 \times 3 \times 3$ supercells, 7 Va, $x_{\mathrm{Gd}} = 0.1296$, PBEsol exchange-correlation functional, $a = 5.402$ Å) and compare CIs obtained from its subsets of 73 and 163 structures with CIs obtained from the full set (210 structures). The results are shown in Figs. S5 and S6. CIs obtained from the subset of 73 structures shows good agreement with CIs obtained from 210 structures, while CIs obtained from the subset of 163 structures shows even better agreement. Thus CIs converge with the increase of number of structures and the larger set provides more accurate CIs. We conclude that for this particular case ($3 \times 3 \times 3$ supercells, 32 unknown CIs) a set of ca 150 structures is large enough. Also, one should care about a composition of the set of structures, so that each CI can be found in at least several number of structures.

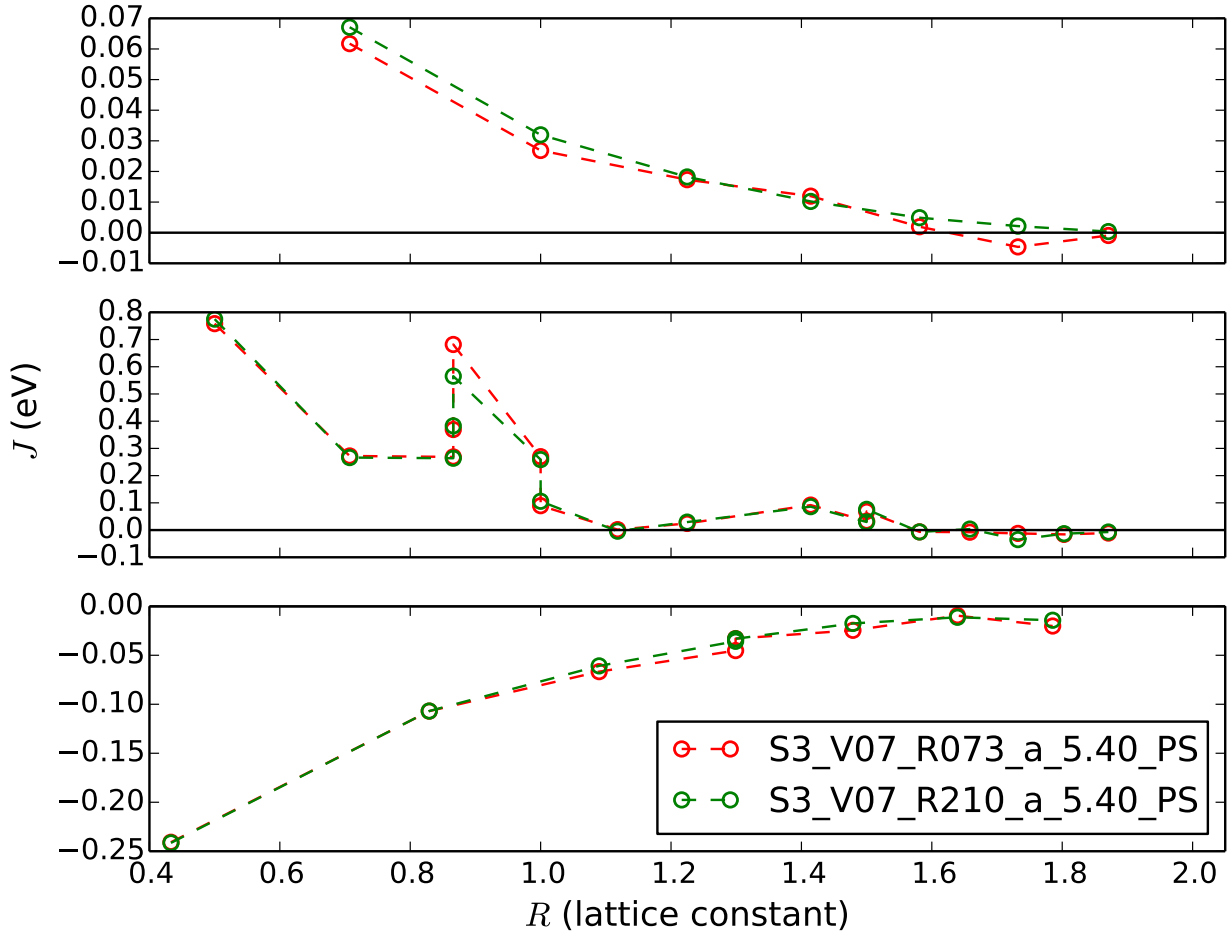

Figure S5: Comparison of CIs obtained from 210 structures and a subset of 73 structures. CIs (J's) *vs.* distance, from top to bottom: Gd−Gd, Va−Va, Gd−Va.

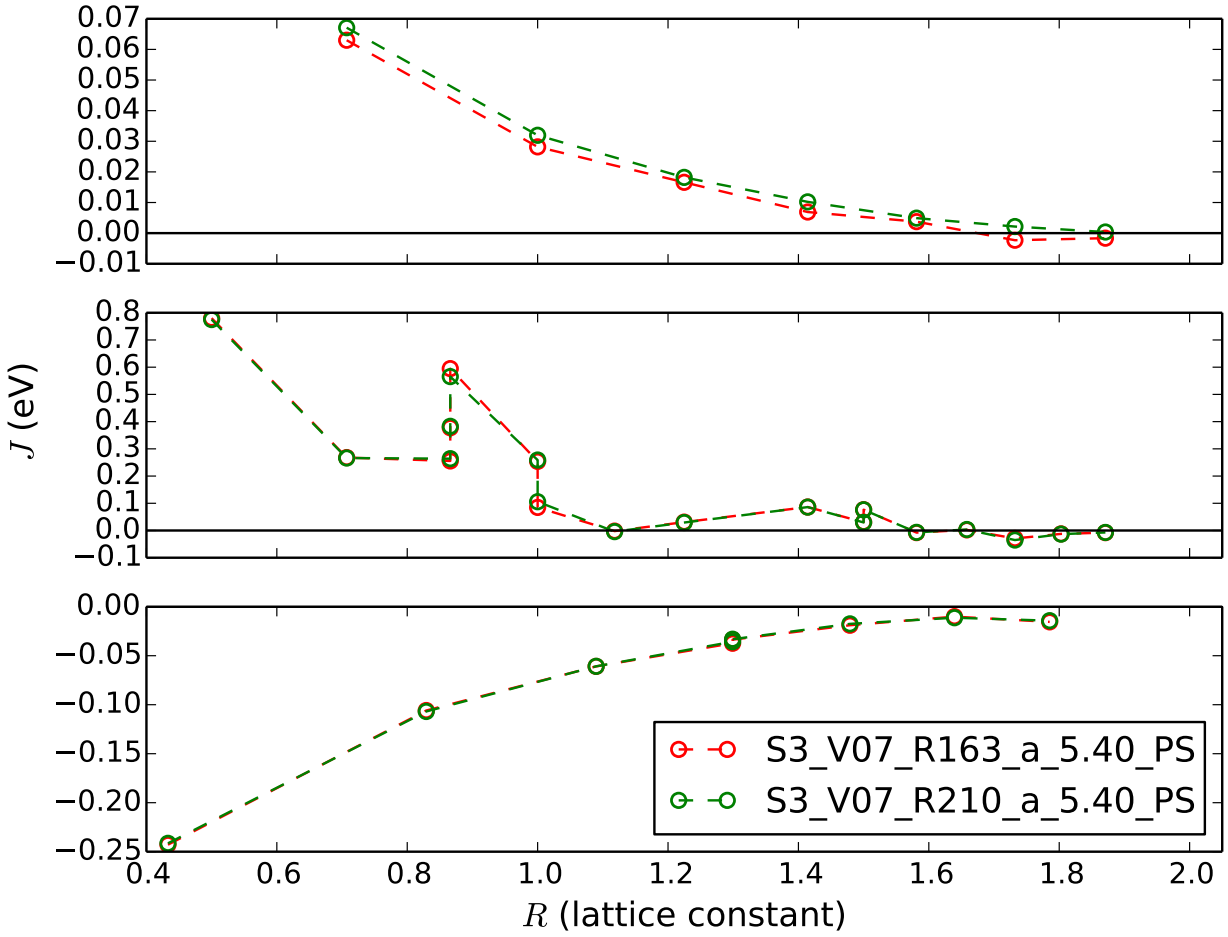

Figure S6: Comparison of CIs obtained from 210 structures and a subset of 163 structures. CIs (J's) *vs.* distance, from top to bottom: Gd−Gd, Va−Va, Gd−Va.

1.e Effect of exchange-correlation functional on CIs

From Figs. S7 and S8 one can see that the choice of exchange-correlation functional (within Generalized Gradient Approximation) does not affect significantly CIs. For $a = 5.402$ Å and $a = 5.470$ Å, both PBEsol and PBE (in Figs. S7 and S8 we call them PS and PE, respectively) provide very similar CIs.

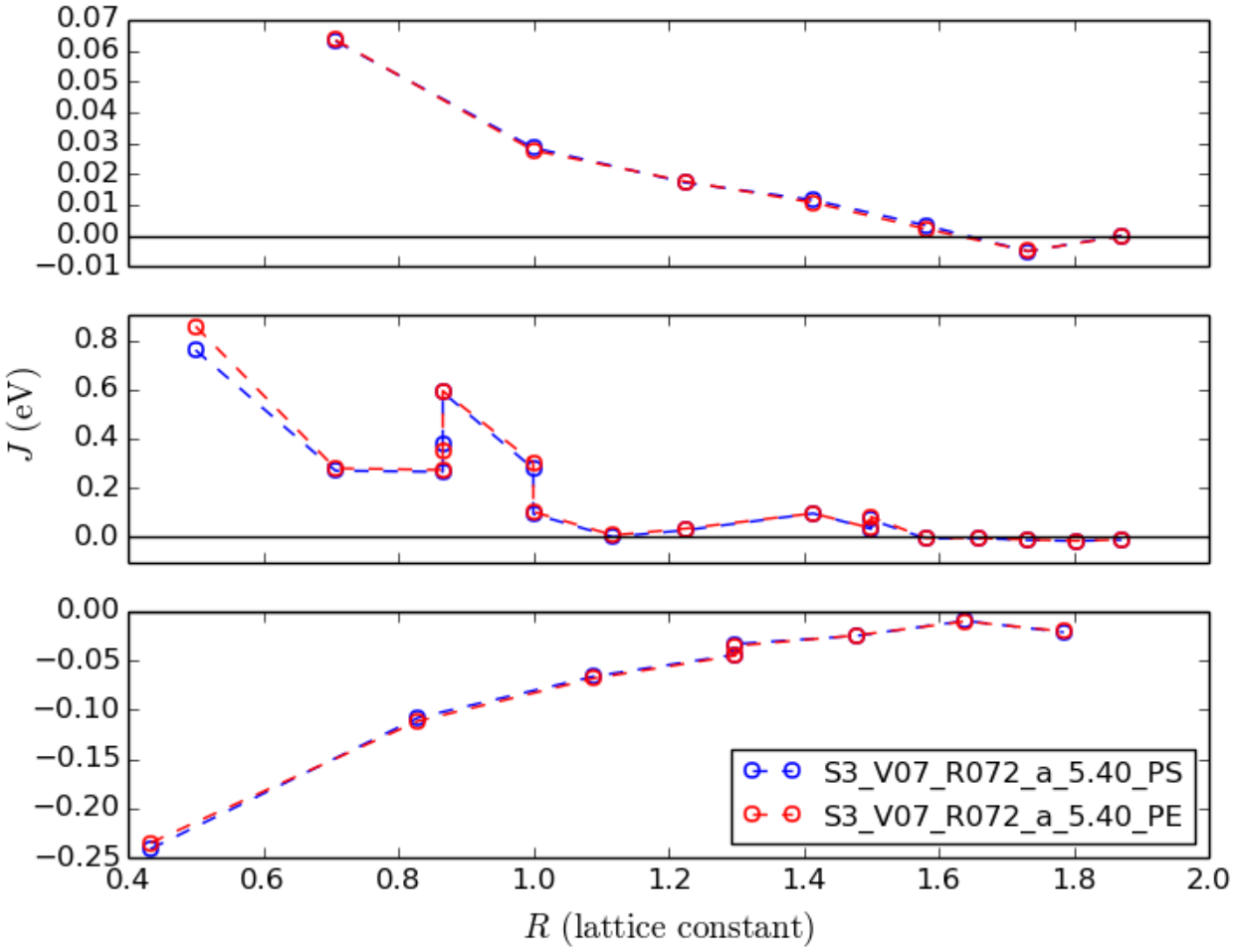

Figure S7: Comparison of CIs obtained from the same set of structures using PBEsol and PBE exchange-correlation functionals. $a = 5.402$ Å. CIs (J's) *vs.* distance, from top to bottom: Gd−Gd, Va−Va, Gd−Va.

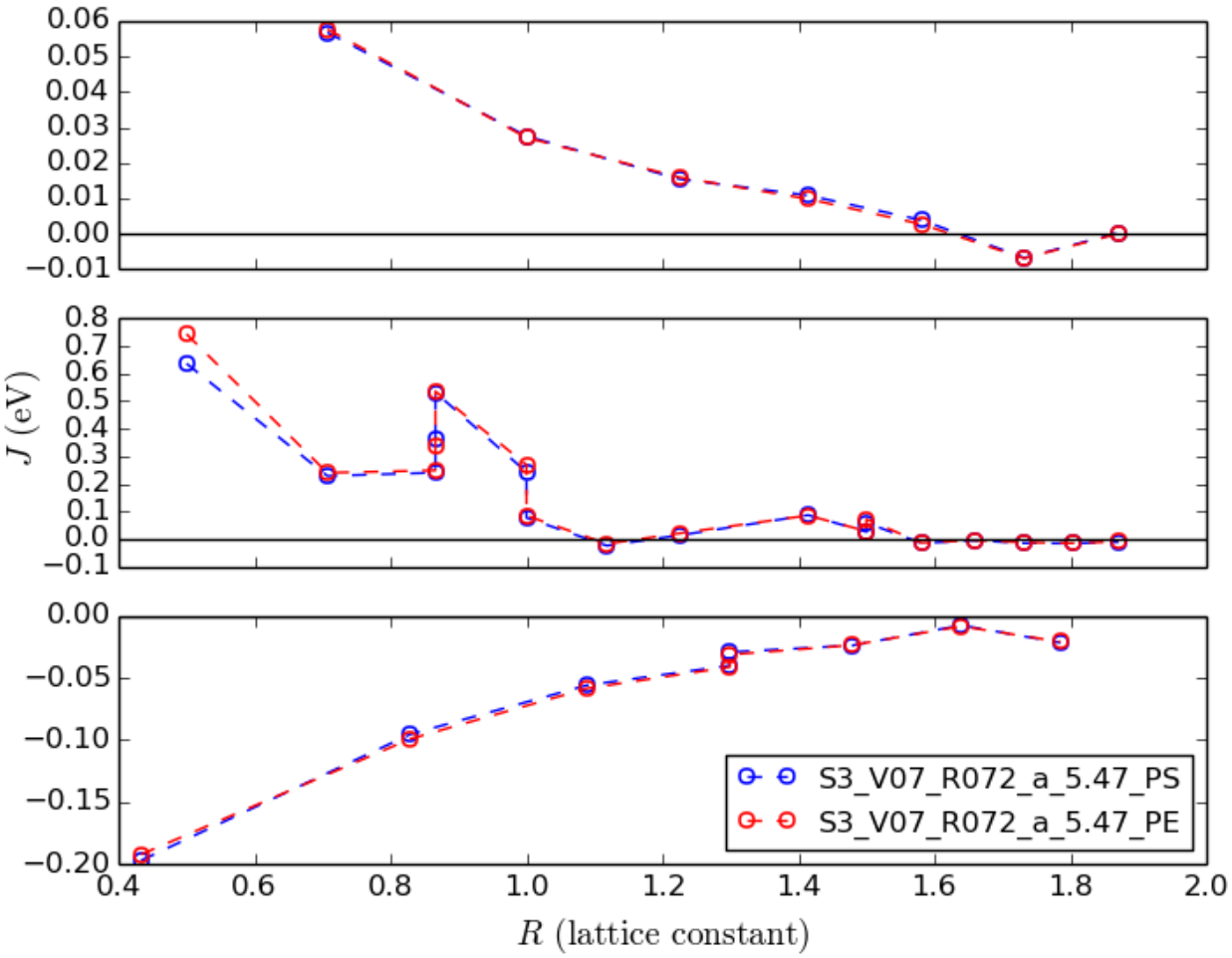

Figure S8: Comparison of CIs obtained from the same set of structures using PBEsol and PBE exchange-correlation functionals. $a = 5.470$ Å. CIs (J's) *vs.* distance, from top to bottom: Gd−Gd, Va−Va, Gd−Va.

1.f Effect of volume expansion on CIs

From Figs. S9 and S10 one can see that volume expansion does not affect the shape of CIs curves, but only scales them (absolute values become smaller).

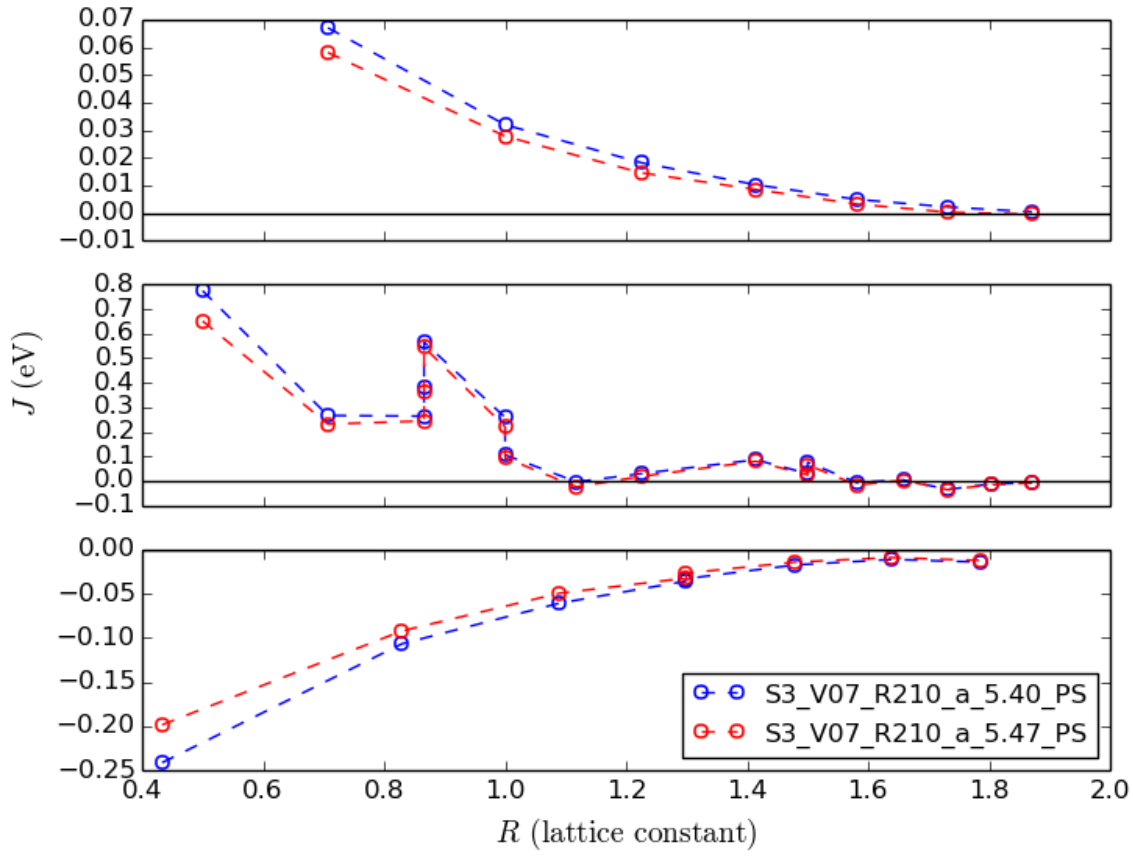

Figure S9: Comparison of CIs obtained from two sets with the same structures but different lattice parameters. $3 \times 3 \times 3$ supercells. CIs (J's) *vs.* distance from top to bottom: Gd−Gd, Va−Va, Gd−Va.

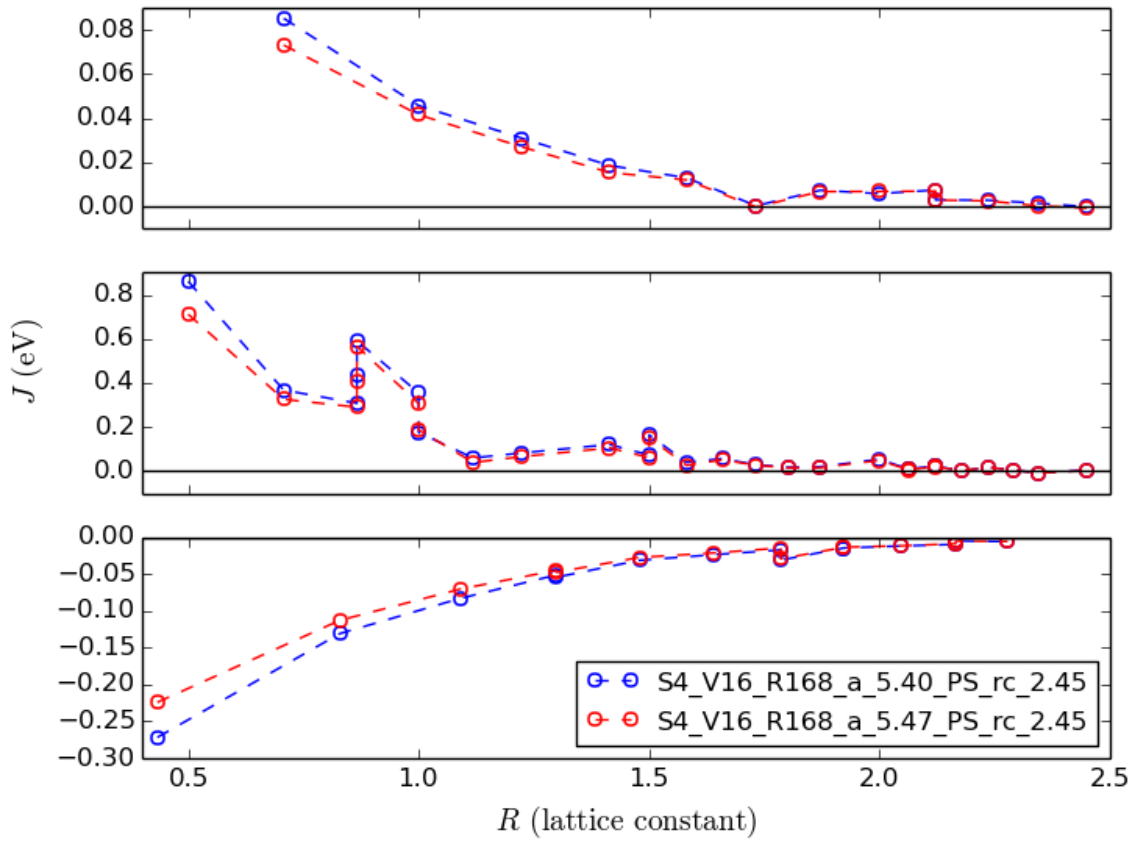

Figure S10: Comparison of CIs obtained from two sets with the same structures but different lattice parameters. $4 \times 4 \times 4$ supercells. CIs (J's) *vs.* distance from top to bottom: Gd−Gd, Va−Va, Gd−Va.

1.g Concentration dependence of CIs

In order to understand how CIs vary with Gd concentration we have performed DFT calculations for structures with different x_{Gd} and performed cluster expansion for each set separately (see the main part of this paper). Here we show that within the $0.09 < x_{\mathrm{Gd}} < 0.25$ range CIs are very similar (see Fig. S11).

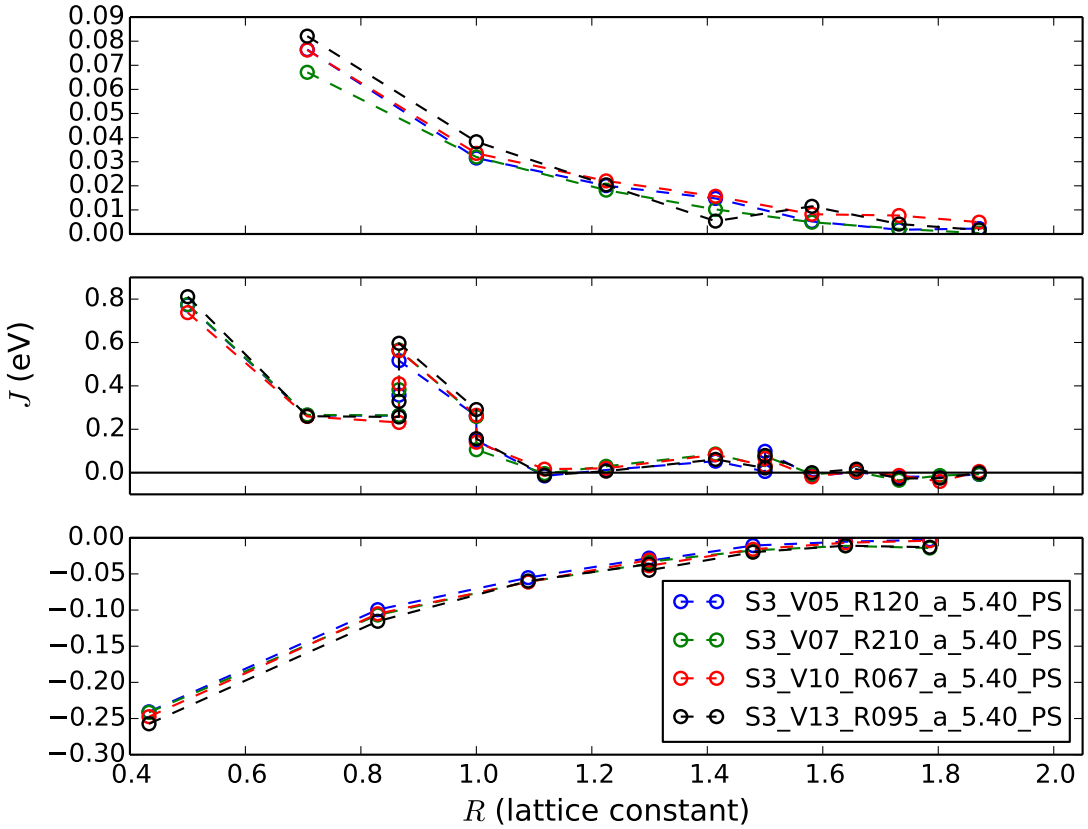

Figure S11: Comparison of CIs obtained for different Gd concentrations in the range $0.09 < x_{\mathrm{Gd}} < 0.25$ ("V05": $x_{\mathrm{Gd}} = 0.0926$, "V07": $x_{\mathrm{Gd}} = 0.1296$, "V10": $x_{\mathrm{Gd}} = 0.1852$, "V13": $x_{\mathrm{Gd}} = 0.2407$). CIs (J's) *vs.* distance from top to bottom: Gd−Gd, Va−Va, Gd−Va.

Also in Fig. S12 we present the two–set–CV scores[2], when making cross–predictions using these CIs (see Fig. S11). In the worst case the two–set–CV score is 7.3 meV/defect ($N_{\mathrm{defects}} = 39$), or 0.9 meV/lattice site.

It also follows from Fig. S12 that the sets with larger number of structures provide CIs with better predictive capability. For example, the predictions with larger two–set–CV scores rely on the CIs calculated using the sets with $N_{\mathrm{struct.}} < 100$, while the set with maximal number of structures (210) shows the best performance. For the case of GDC and supercell size ($3 \times 3 \times 3$) with 32 unknown CIs in CE, we suggest to have a $N_{\mathrm{struct.}} \gtrsim 120$ in order to obtain reliable CIs.

[2]Additionally, to characterise the predictive capabilities of our CE we calculate a two–set–cross validation (two–set–CV) score, which is similar to the "self" CV score, but $E_i^{'}$ are calculated using CIs obtained from a different set of structures. Since configurational energy is defined only up to an additive constant, in this case we naturally apply a constant shift to all $E_i^{'}$ in order to minimise the two-set-CV score.

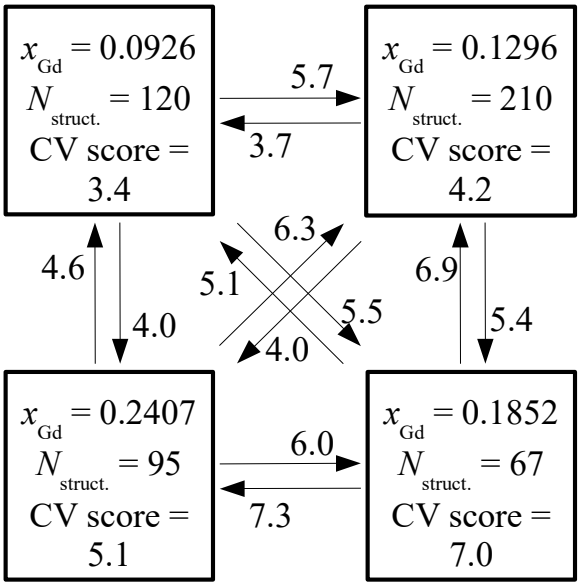

Figure S12: Each box represents the set of DFT structures, showing Gd concentration, number of structures and CV score. The following four sets are used: set_02, set_03, set_04, set_05 (see Appendix 1). All possible twelve two-set-CV scores (meV/defect) for these four sets are shown with arrows. For example, the two-set-CV score equals 3.7 meV/defect when predicting energies of the set with $x_{\mathrm{Gd}} = 0.0926$, while using the CIs obtained from the set with $x_{\mathrm{Gd}} = 0.1296$ (arrow from the right-upper box to the left-upper box). Normalisation constant is the number of defects in structures, whose energies are predicted.

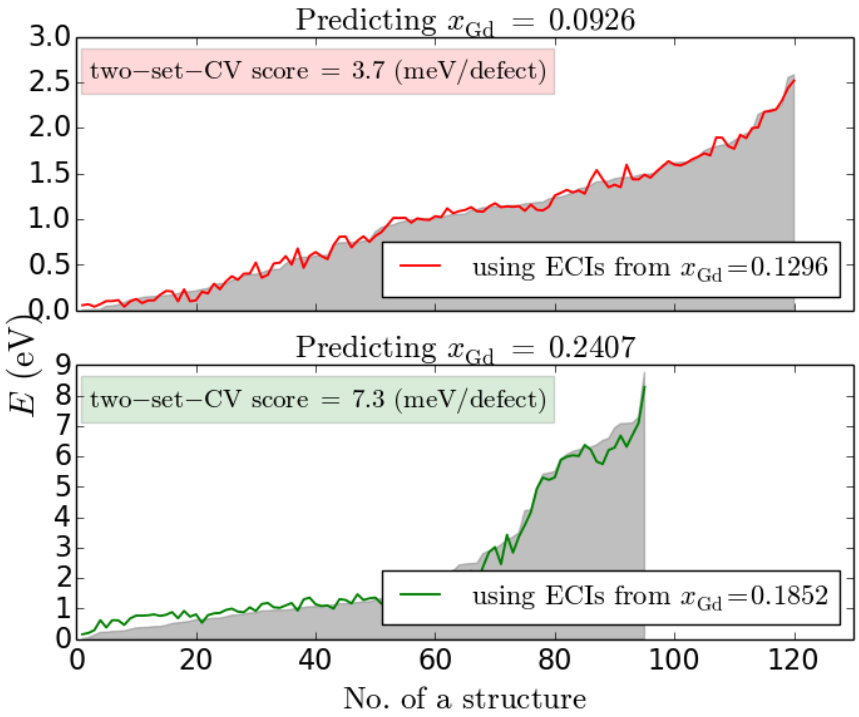

Figure S13: The best (top panel) and the worst (bottom panel) cross-concentration predictions: CE energies (solid lines) are compared with DFT calculated (grey filling), and are aligned to have best fits (Hamiltonian is defined up to an additive constant). Total energies (E) are given (not normalised per number of defects).

Finally, in order to demonstrate how the energies of the structures obtained with our cluster expansion compare to those calculated with DFT, we show the best and the worst cross-concentration predictions (see Fig. S13). The best one yields 3.7 meV/defect (0.2 meV/lattice site) for the prediction of $x_{Gd} = 0.0926$ using CIs obtained for $x_{Gd} = 0.1296$, and the worst one yields 7.3 meV/defect (0.9 meV/lattice site) for the prediction of $x_{Gd} = 0.2407$ using CIs extracted for $x_{Gd} = 0.1852$.

1.h Effect of supercell size on CIs

The supercell size determines how many long-range interactions can be found. This is so due to the adopted lattice model (in CE) and periodic boundary conditions. For example, in any $2\times2\times2$ supercell the number of Gd−Gd CIs with vector $\left\langle \frac{1}{2}, \frac{1}{2}, 0\right\rangle$ equals to the number of Gd−Gd CIs with vector $\left\langle \frac{3}{2}, \frac{1}{2}, 0\right\rangle$, therefore one has to exclude the latter CIs (and more long-range) from CE (see also Appendix 2 in the main paper).

Here we compare CIs obtained from three sets: 1) $2\times2\times2$ large supercells, 206 structures ($x_{\mathrm{Gd}} = 0.0625, 0.1250, 0.1875$); 2) $3\times3\times3$ large supercells, 210 structures ($x_{\mathrm{Gd}} = 0.1296$); 3) $4\times4\times4$ large supercells, 168 structures ($x_{\mathrm{Gd}} = 0.1250$). $a = 5.402$ Å for all sets.

Let us notice that all these sets have different cut-off radii due to different supercell sizes. Therefore, we first of all compare the CIs obtained using different cut-off radii, but using only one fixed supercell size ($4\times4\times4$ supercells). Interactions with $R \leq R^{\mathrm{CE}}_{\mathrm{CUT}}$ are included:

a) $R^{\mathrm{CE}}_{\mathrm{CUT}} = 2.45a$ for Gd−Gd and Va−Va CIs, and $R^{\mathrm{CE}}_{\mathrm{CUT}} = 2.278a$ for Gd−Va (we also exclude CIs with the vector $\left\langle \frac{9}{4}, \frac{1}{4}, \frac{1}{4}\right\rangle$);

b) $R^{\mathrm{CE}}_{\mathrm{CUT}} = 2.0a$;

c) like in the case of S3 large supercells: $R^{\mathrm{CE}}_{\mathrm{CUT}} = 1.871a$ for all CIs, except $1.786a$ for Gd−Va and we also exclude Gd−Va vector $\left\langle \frac{7}{4}, \frac{1}{4}, \frac{1}{4}\right\rangle$ ('cut like S3' case);

d) like in the case of S2 large supercells: $R^{\mathrm{CE}}_{\mathrm{CUT}} = 1.5a$ for all CIs, except $1.3a$ for Gd−Va and we also exclude Gd−Va vector $\left\langle \frac{5}{4}, \frac{1}{4}, \frac{1}{4}\right\rangle$ and Va−Va vector $\left\langle \frac{3}{2}, 0, 0\right\rangle$ ('cut like S2' case).

The results are shown in Fig. S14. One can see that as the cut-off radius decreases CIs shift towards $J = 0$ (absolute values become smaller while the shape of CIs curves remain). This is so because the long-range CIs are cut and the less long-range CIs effectively become a zero.

Now let us compare CIs obtained from above-mentioned sets (different supercell sizes) using corresponding cut-off radii. Fig. S15 shows that for smaller supercell sizes CIs are shifted towards $J = 0$ (again, less long-range interactions effectively become zero).

Thus both smaller supercells and shorter cut-off radii cause the shift of CIs towards zero. This reflects the electrostatic nature of these interactions. Keeping in mind that CIs are calculated with respect to "effective zero" ($J = 0$), we would like to mention the following about the effect of supercell size on CIs:

1) Absolute values of long-range ($\gtrsim 8$ Å) CIs calculated within the CE method in the static limit including ionic relaxations are relatively small (order of magnitude is 1–60 meV).

2) One should recall that the concept of a cluster interaction (CI) is an approximation, attempting to *average individual interactions*. In general, *individual interactions* are affected by a local structure, which should cause a scatter in values of *individual interactions*. The *averaging* is valid only if the scatter is modest. One could expect that in the case when interactions are close to zero, the scatter can easily spoil the concept of CI (it would be meaningless to, *e.g.* average $J_1 = j$ and $J_2 = -j$ and obtain $J_{\mathrm{average}} = 0$). Nevertheless, our test calculations show that the concept of CI is reasonable for the considered here system (see error bars in Fig. S16).

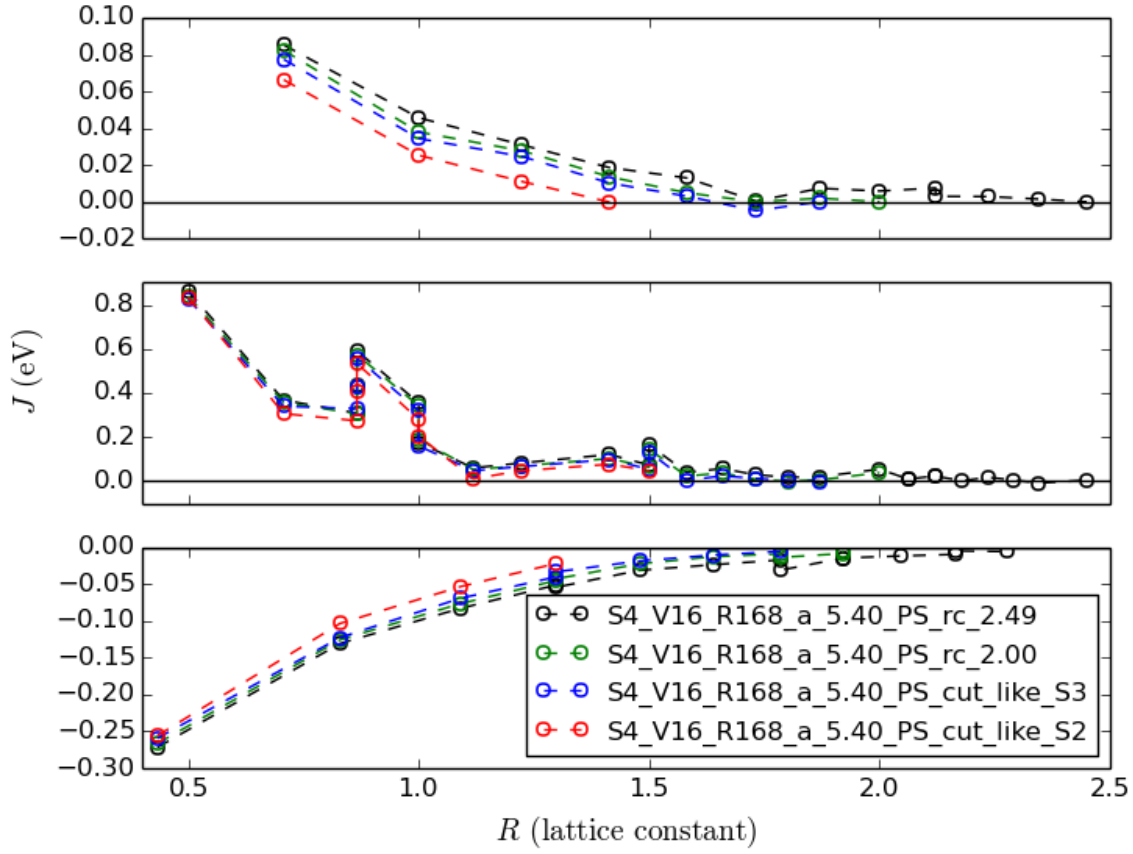

Figure S14: Comparison of CIs obtained from $4 \times 4 \times 4$ large supercells using different cut-off radii (rc means cut-off radius, while cut_like means cut-off radius like one used for smaller supercells; see text for details). CIs (J's) *vs.* distance, from top to bottom: Gd−Gd, Va−Va, Gd−Va.

3) From simple electrostatic reasoning Va−Va CIs should be repulsive (positive J). This behaviour for long-range Va−Va CIs is not well reproduced using $3 \times 3 \times 3$ supercells. However, it is reproduced using $4 \times 4 \times 4$ supercells (see Fig. S16). Therefore, the $4 \times 4 \times 4$ supercell is the first supercell to show qualitatively reasonable results (positioning of long-range Va−Va CIs relative to effective zero).

4) Let us notice that the calculation of larger supercells with DFT is not feasible (*e.g.* S5 supercell contains 1500 lattice sites). We expect that for larger supercells one would obtain CIs shifted in comparison to S4. Notwithstanding, the absolute values of our CIs are close to those obtained by Grieshammer *et al.*[3], where finite size correction was applied (extrapolation to infinite supercell size).

5) In our case, the application of a finite size correction to CIs (due to the finite supercell size and long-range nature of the electrostatic interactions) is not well justified as while resolving one ambiguity — correcting the absolute values of CIs, it might introduce another one — a

[3]S. Grieshammer, B. O. H. Grope, J. Koettgen and M. Martin, 'A combined DFT + U and Monte Carlo study on rare earth doped ceria', *Phys. Chem. Chem. Phys.*, **16**, 9974-9986, 2014.

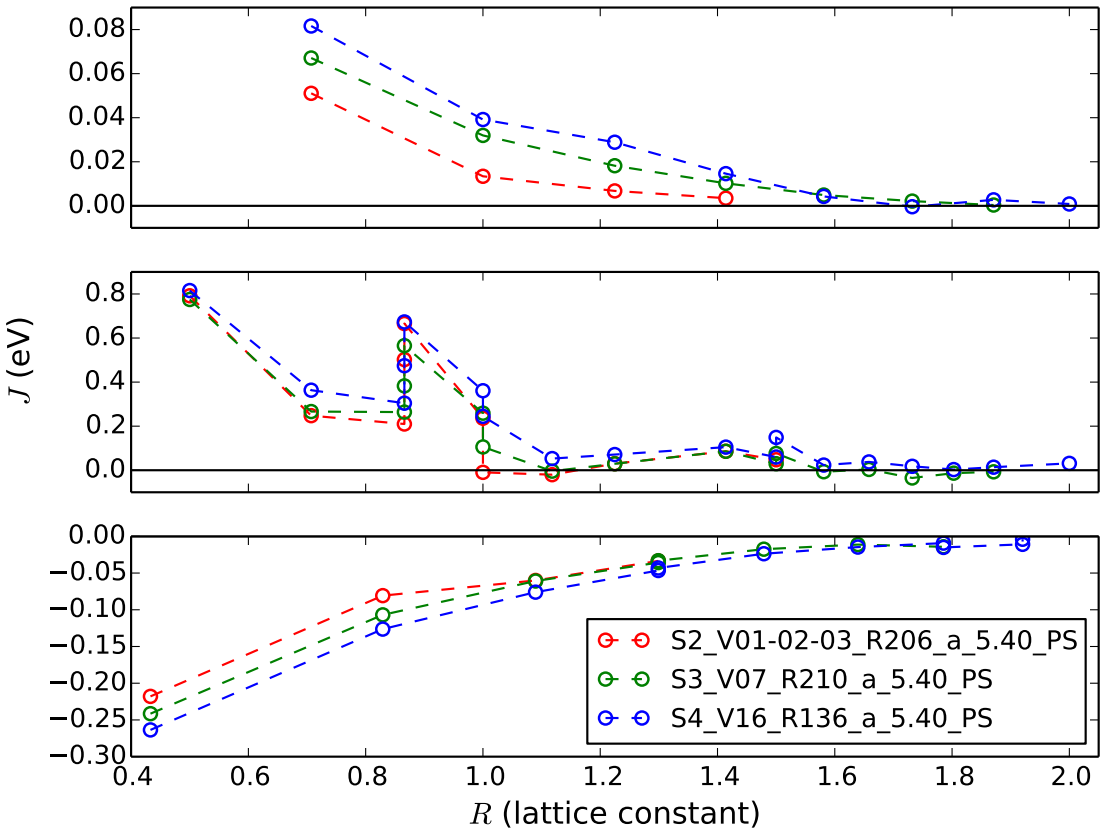

Figure S15: Comparison of CIs obtained from $2 \times 2 \times 2$, $3 \times 3 \times 3$ and $4 \times 4 \times 4$ large supercells. CIs (J's) *vs.* distance from top to bottom: Gd−Gd, Va−Va, Gd−Va.

change of the shape of the potential (a shift of previously effectively zero interactions without determination of the tail).

Taking into account these arguments and Monte Carlo test simulations we further demonstrate that CIs obtained using S4 supercells are reliable for Monte Carlo simulations (see Section 2 for further discussions regarding Monte Carlo simulations).

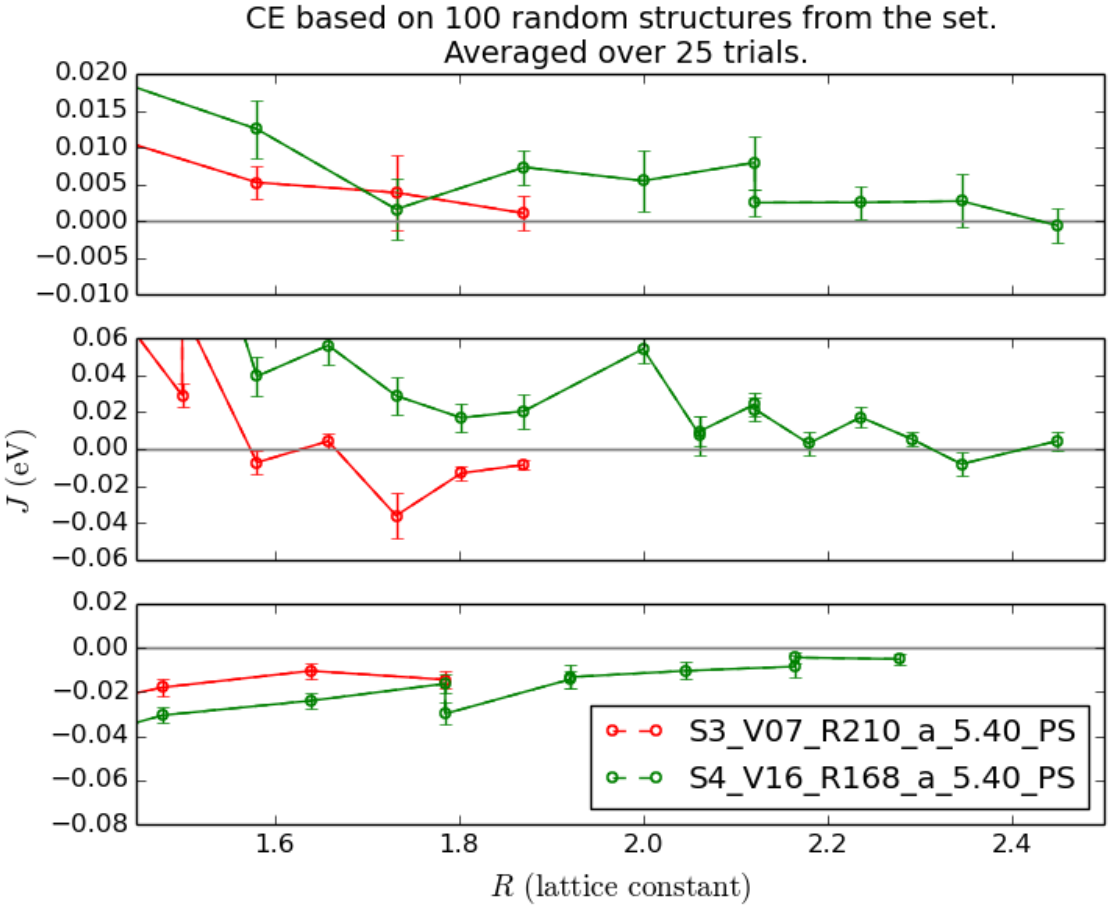

Figure S16: Comparison of long-range CIs obtained from $3 \times 3 \times 3$ (210 structures) and $4 \times 4 \times 4$ (168 structures) large supercells. CIs were averaged over 25 cluster expansions, each was based on 100 random structures from the set. Errorbars show standard deviations. CIs (J's) *vs.* distance, from top to bottom: Gd−Gd, Va−Va, Gd−Va.

1.i CIs in the case of no ionic relaxation

In the case of no relaxation, CIs can be fitted with the $C + \frac{k}{R}e^{-R/\lambda}$ (see Fig. S17).

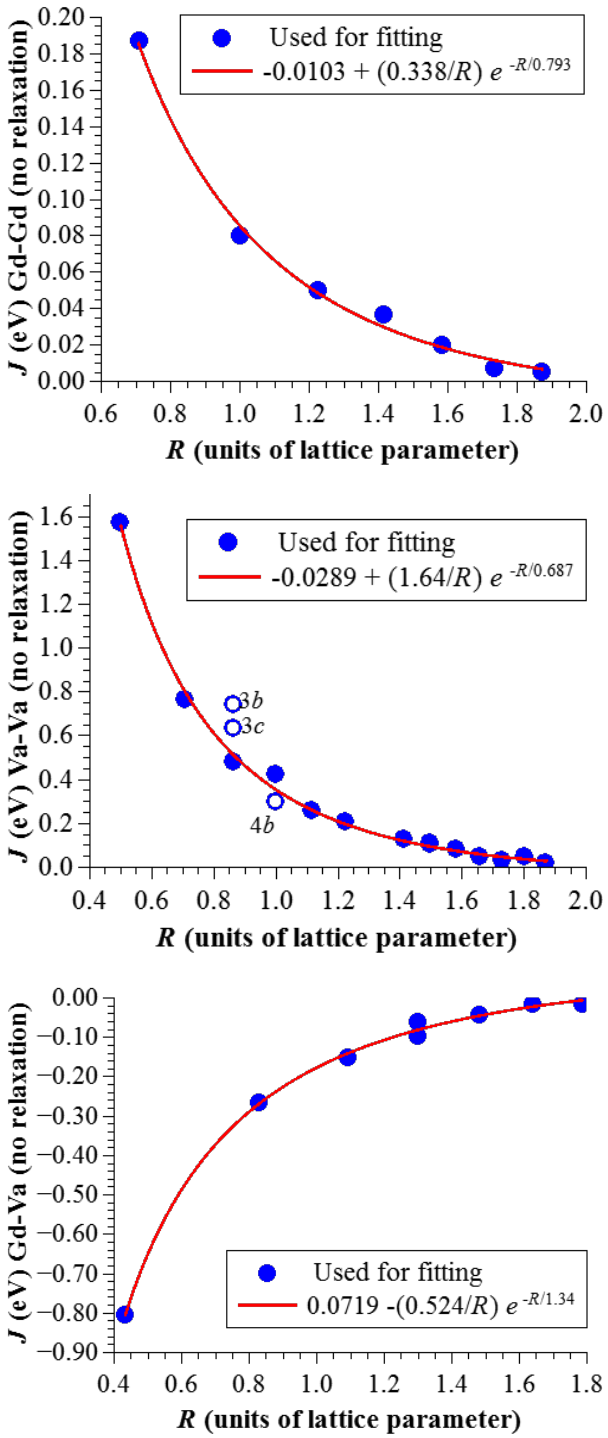

Figure S17: Gd−Gd, Va−Va and Gd−Va CIs *vs.* distance calculated from 210 structures (set S3_V07_N210_a_540_PS, $x_{\mathrm{Gd}} = 0.1296$) with no ionic relaxation. Red lines show fits of pair CIs with $C + \frac{k}{R}e^{-R/\lambda}$. CIs not included in the fit are labelled in accordance with the Table 2 (see the main part of this paper).

1.j List of CIs

Here we list the CIs obtained using set S4_V16_R168_a_5.47_PS (set_12 in Appendix 1 in the main part of this paper) and $\mathcal{H}_5$ (see Tables S4, S5 and S6).

Table S4: Gd−Gd CIs calculated from the set S4_V16_R168_a_5.47_PS (set_12, see Appendix 1 in the main part of this paper). Vector $\vec{r}$ connects interacting Gd−Gd on fcc Ce&Gd sublattice. $|\vec{r}|$ is the length of $\vec{r}$, and $\vec{r}_x$, $\vec{r}_y$, $\vec{r}_z$ — its components. Vectors are given in the units of CeO_2 conventional cell (*cF12* in Pearson notation).

$\lvert\vec{r}\rvert$	$\vec{r}_x$	$\vec{r}_y$	$\vec{r}_z$	CI (eV)
0.70711	0.5	0.5	0.0	0.07311
1.00000	1.0	0.0	0.0	0.04154
1.22474	1.0	0.5	0.5	0.02693
1.41421	1.0	1.0	0.0	0.01536
1.58114	1.5	0.5	0.0	0.01201
1.73205	1.0	1.0	1.0	-0.00027
1.87083	1.5	1.0	0.5	0.00672
2.00000	2.0	0.0	0.0	0.00677
2.12132	1.5	1.5	0.0	0.00695
2.12132	2.0	0.5	0.5	0.00288
2.23607	2.0	1.0	0.0	0.00250
2.34521	1.5	1.5	1.0	0.00022
2.44949	2.0	1.0	1.0	-0.00069

Table S5: Va−Va CIs calculated from the set S4_V16_R168_a_5.47_PS (set_12, see Appendix 1 in the main part of this paper). Vector $\vec{r}$ connects interacting Va−Va on simple cubic O&Va sublattice. $|\vec{r}|$ is the length of $\vec{r}$, and $\vec{r}_x$, $\vec{r}_y$, $\vec{r}_z$ — its components. Vectors are given in the units of CeO_2 conventional cell (*cF12* in Pearson notation). Column 'Comment' refers to the CIs labeling introduced in the main part of this paper (in order to avoid ambiguities).

$\lvert\vec{r}\rvert$	$\vec{r}_x$	$\vec{r}_y$	$\vec{r}_z$	Comment	J (eV)
0.50000	0.5	0.0	0.0	-	0.71145
0.70711	0.5	0.5	0.0	-	0.32594
0.86603	0.5	0.5	0.5	3a	0.29047
0.86603	0.5	0.5	0.5	3b	0.40885
0.86603	0.5	0.5	0.5	3c	0.56556
1.00000	1.0	0.0	0.0	4a	0.30724
1.00000	1.0	0.0	0.0	4b	0.18797
1.11803	1.0	0.5	0.0	-	0.03991
1.22474	1.0	0.5	0.5	-	0.06657
1.41421	1.0	1.0	0.0	-	0.10573
1.50000	1.0	1.0	0.5	-	0.06217
1.50000	1.5	0.0	0.0	-	0.15172
1.58114	1.5	0.5	0.0	-	0.02664
1.65831	1.5	0.5	0.5	-	0.05185
1.73205	1.0	1.0	1.0	-	0.02738
1.80278	1.5	1.0	0.0	-	0.01542
1.87083	1.5	1.0	0.5	-	0.01631
2.00000	2.0	0.0	0.0	-	0.04925
2.06155	1.5	1.0	1.0	-	0.00743
2.06155	2.0	0.5	0.0	-	0.00609
2.12132	1.5	1.5	0.0	-	0.02144
2.12132	2.0	0.5	0.5	-	0.01960
2.17945	1.5	1.5	0.5	-	0.00344
2.23607	2.0	1.0	0.0	-	0.01777
2.29129	2.0	1.0	0.5	-	0.00424
2.34521	1.5	1.5	1.0	-	-0.00806
2.44949	2.0	1.0	1.0	-	0.00395

Table S6: Gd−Va CIs calculated from the set S4_V16_R168_a_5.47_PS (set_12, see Appendix 1 in the main part of this paper). Vector $\vec{r}$ connects interacting Gd−Va going from Ce&Gd sublattice to O&Va sublattice. $|\vec{r}|$ is the length of $\vec{r}$, and $\vec{r}_x$, $\vec{r}_y$, $\vec{r}_z$ — its components. Vectors are given in the units of CeO_2 conventional cell (*cF12* in Pearson notation).

$\lvert\vec{r}\rvert$	$\vec{r}_x$	$\vec{r}_y$	$\vec{r}_z$	J (eV)
0.43301	0.25	0.25	0.25	-0.22400
0.82916	0.75	0.25	0.25	-0.11254
1.08972	0.75	0.75	0.25	-0.07022
1.29904	0.75	0.75	0.75	-0.04455
1.29904	1.25	0.25	0.25	-0.04722
1.47902	1.25	0.75	0.25	-0.02706
1.63936	1.25	0.75	0.75	-0.02110
1.78536	1.25	1.25	0.25	-0.01420
1.78536	1.75	0.25	0.25	-0.02904
1.92029	1.25	1.25	0.75	-0.01359
1.92029	1.75	0.75	0.25	-0.01323
2.04634	1.75	0.75	0.75	-0.01130
2.16506	1.25	1.25	1.25	-0.00853
2.16506	1.75	1.25	0.25	-0.00464
2.27761	1.75	1.25	0.75	-0.00525

1.k Effect of spatial distribution of Gd and Va on electronic structure and ionic charges

The above-mentioned list of CIs (Section 1.j) can be used to describe the energetics of configurations with different degrees of ordering&clustering encountered in this study, as the underlying electronic structure remains fairly the same. Let us consider two configurations in $4 \times 4 \times 4$ supercells – one quite random and another with Gd and Va clustered together (4 eV lower in energy). In Fig. S18 we show the Density of states (DOS) for both supercells. Ordering and clustering yields sharpening of DOS peaks (bottom panel in Fig. S18) as compared to random-like distribution (top panel in Fig. S18), but the shape of DOS and, most important, energy gaps are very much the same. Similarly, fluctuations of Bader charges due to the different distribution of ions are estimated to be less than $0.15e$ for O^{2-} and ca three times smaller for cations. Similarly, fluctuations of Bader charges due to the different distribution of ions are estimated to be less than $0.15e$ for O^{2-} and ca three times smaller for cations.

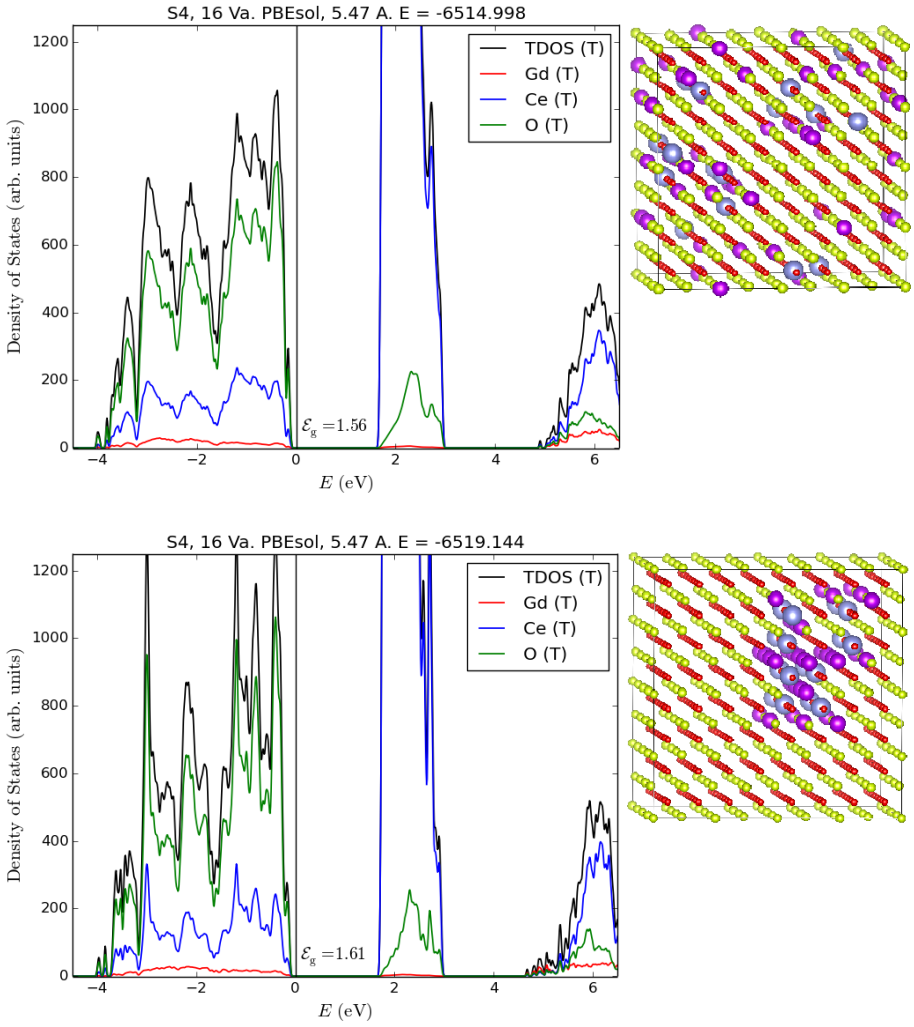

Figure S18: Density of states (DOS) of two configurations ($4 \times 4 \times 4$ supercells). Top panel: random-like distribution of Gd and Va. Bottom panel: clustering of Gd and Va resembling phase separation. E_g is the gap between valence band maximum and unoccupied Ce 4f states. Colour code: yellow – Ce, violet – Gd, red – O, grey – Va.

2 Supplementary information to the section *Monte Carlo simulations*

2.a MC simulations using both different CE cut-off radii and MC cut-off radii, and different supercell sizes (DFT&CE)

When performing the cluster expansion we use a certain cut-off radius[4] $R_{\mathrm{CUT}}^{\mathrm{CE}}$. When performing Monte Carlo simulations we use another cut-off radius $R_{\mathrm{CUT}}^{\mathrm{MC}}$.

Here we would like to consider set S4_V16_R168_PS_a_5.40 and CIs obtained from it using different $R_{\mathrm{CUT}}^{\mathrm{CE}}$: $2.45a$, $2.00a$, 'cut like S3' and 'cut like S2' (see Fig. S14). For each of these four sets of CIs we use a $R_{\mathrm{CUT}}^{\mathrm{MC}}$ going from $1.00a$ to $2.45a$. We also consider set S4_V16_R168_PS_a_5.47 ($R_{\mathrm{CUT}}^{\mathrm{CE}} = 2.45a$), set S3_V07_R210_PS_a_5.40 ('cut like S3') and set S2_V01-V02-V03_R206_PS_a_5.40 ('cut like S2'). The O−Va short range order (SRO) parameters (α_i) for different coordination shells are shown in Figs. S19, S20, S21, S22, S23, S24, S25. These serve as fingerprints for the whole structure, therefore, we do not show here the SRO parameters for CeGd and GdVa.

Regarding the S4 large supercells, from Figs. S19, S20, S21, S22, S23 we conclude the following:

1) Most accurate simulations show that the ground state solution is the decomposition into CeO_2 and C-type Gd_2O_3. We further call this solution the ground state (GS).

2) If small $R_{\mathrm{CUT}}^{\mathrm{CE}}$ is used ("cut like S2"), Monte Carlo simulations converge to a wrong GS.

3) For $R_{\mathrm{CUT}}^{\mathrm{MC}} \lesssim 1.45a$ we find no convergence to the GS.

4) In the interval $1.45a \lesssim R_{\mathrm{CUT}}^{\mathrm{MC}} \lesssim 2.1a$ apart from the GS solution we also find other solutions.

5) After $R_{\mathrm{CUT}}^{\mathrm{MC}} \gtrsim 2.1a$ GS solution is stabilised (see, *e.g.* the tail plateau in Fig. S22). To the best of our knowledge, $R_{\mathrm{CUT}}^{\mathrm{CE}} = 2.45a$ and $R_{\mathrm{CUT}}^{\mathrm{MC}} \gtrsim 2.1a$ guarantee the convergence to the GS (for both lattice parameters, namely $a = 5.402, 5.470$ Å).

For comparison we also show the same MC test with the CIs calculated from set S3_V07_R210_PS_a_5.40 (Fig. S24). The GS solution is remarkably stable for $R_{\mathrm{CUT}}^{\mathrm{MC}} > 1.4a$. This may be attributed to two factors, both originating from the small supercell size, *i.e.* the shift of CIs: 1) the negative Va−Va CI corresponding to vector $\langle 1, \frac{1}{2}, 0 \rangle$ (compare this Va−Va CI for S3 and S4 in Fig. S15, middle panel); 2) the artificially negative tail of Va−Va CIs (compare Va−Va CIs for S3 and S4 in Fig. S16, middle panel). Similar MC tests using CIs derived from S2 large supercells (Fig. S25, S2_V01-V02-V03_R206_PS_a_5.40) show no convincing convergence to any solution.

Let us also mention that CIs obtained from the S3 supercells overestimate T_{C} due to the above-mentioned reasons, while CIs obtained from the S4 supercells provide convergent T_{C} for $R_{\mathrm{CUT}}^{\mathrm{MC}} \gtrsim 2.1a$ (see Fig. S26).

Based on the analysis presented in Section 1.h, and the results of our MC tests for the GDC case, we conclude the following:

[4] It might be different for different types of interactions.

1) The cluster expansion should be done using the largest possible supercell size and largest possible cut-off radius (number of structures should be sufficiently large in order to provide CV score minimisation).

2) The CIs derived from S4_V16_R168_a_540_PS and S4_V16_R168_a_547_PS sets ($R_{\mathrm{CE}}^{\mathrm{CUT}} = 2.45a$) are trustworthy, and most precise among calculated. To the best of our knowledge, in Monte Carlo simulations these CIs provide reliable results, converging to the decomposition into CeO_2 and C-type Gd_2O_3 ($R_{\mathrm{MC}}^{\mathrm{CUT}} \gtrsim 2.1a$; for both $a = 5.402$ Å and $a = 5.470$ Å).

3) The CIs derived from smaller supercells and cut-off radii do not guarantee a reliable solution (S2) or artificially over-stabilise the GS solution (S3), thus are not reliable.

Notes on the composition of precipitates

As we already mentioned, MC simulations are rather sensitive to the long-range CIs. Here we briefly summarise the possible outcomes of MC simulations with shorter cut-off radii. Although such outcomes are computational artefacts, they might be considered as "competitive phases". Here we report for the $x_{\mathrm{Gd}} = 0.125$ case.

If not sufficiently large cut-off radius is chosen, the phase separation into CeO_2 and $CeO_2{-}Gd_2O_3$ can occur. It is, in fact, very similar to the phase separation into CeO_2 and Gd_2O_3. The $CeO_2{-}Gd_2O_3$ forms a precipitate, in which the share of Gd_2O_3 is $\approx 60\%$.

The transition temperature and the behaviour of Ce−Gd and Gd−Va SRO parameters are similar to those of pure precipitates (though absolute values of α_i are smaller). The Va−Va clustering in $CeO_2{-}Gd_2O_3$ precipitates depends on the cut-off radius. The O−Va ordering can also differ within the same precipitate. However, the main feature, *i.e.* Va−Va clustering in the $\langle 1, \frac{1}{2}, 0 \rangle$ shells, is always observed. At the same time, the Va−Va clustering in the $\langle \frac{1}{2}, \frac{1}{2}, \frac{1}{2} \rangle$ shells is reduced and even suppressed. Therefore, the Va−Va clustering along $\langle 111 \rangle$, characteristic for pure C-type Gd_2O_3, is not present in such $CeO_2{-}Gd_2O_3$ precipitates.

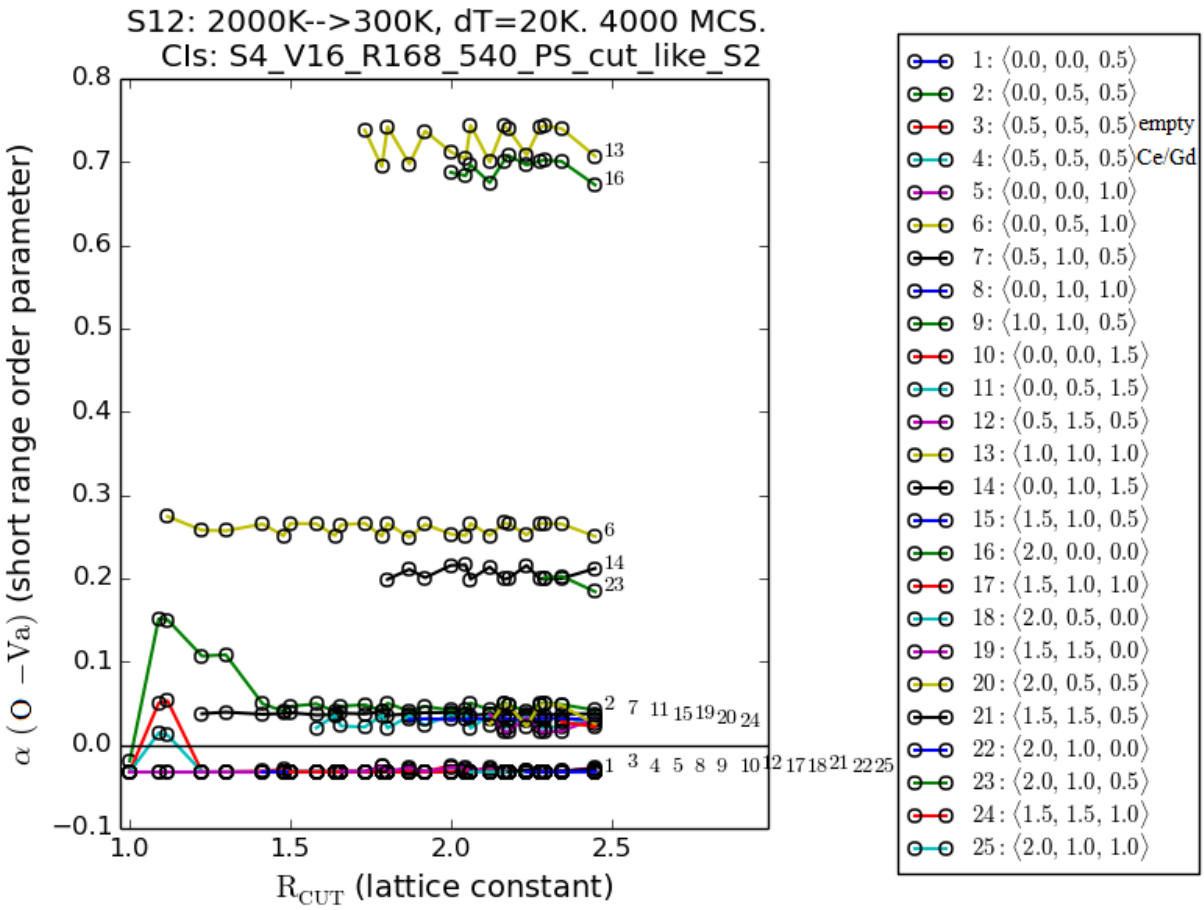

Figure S19: OVa SRO parameters (at 300 K) *vs.* cut-off radius of MC simulation ($x_{\mathrm{Gd}} = 0.125$, $12 \times 12 \times 12$ supercell, annealing from 2000 K to 300 K, 20 K temperature step, 4000 Monte Carlo steps per temperature). CIs obtained using cut-off radius like in the S2 case (thus CIs with $R > R^{\mathrm{CE}}_{\mathrm{CUT}}$ are zeros in MC). Legends show coordination shells (O&Va sublattice). Labels of vectors (shells) are written on the same levels (y coordinates) as the most long-ranged values of α_i's.

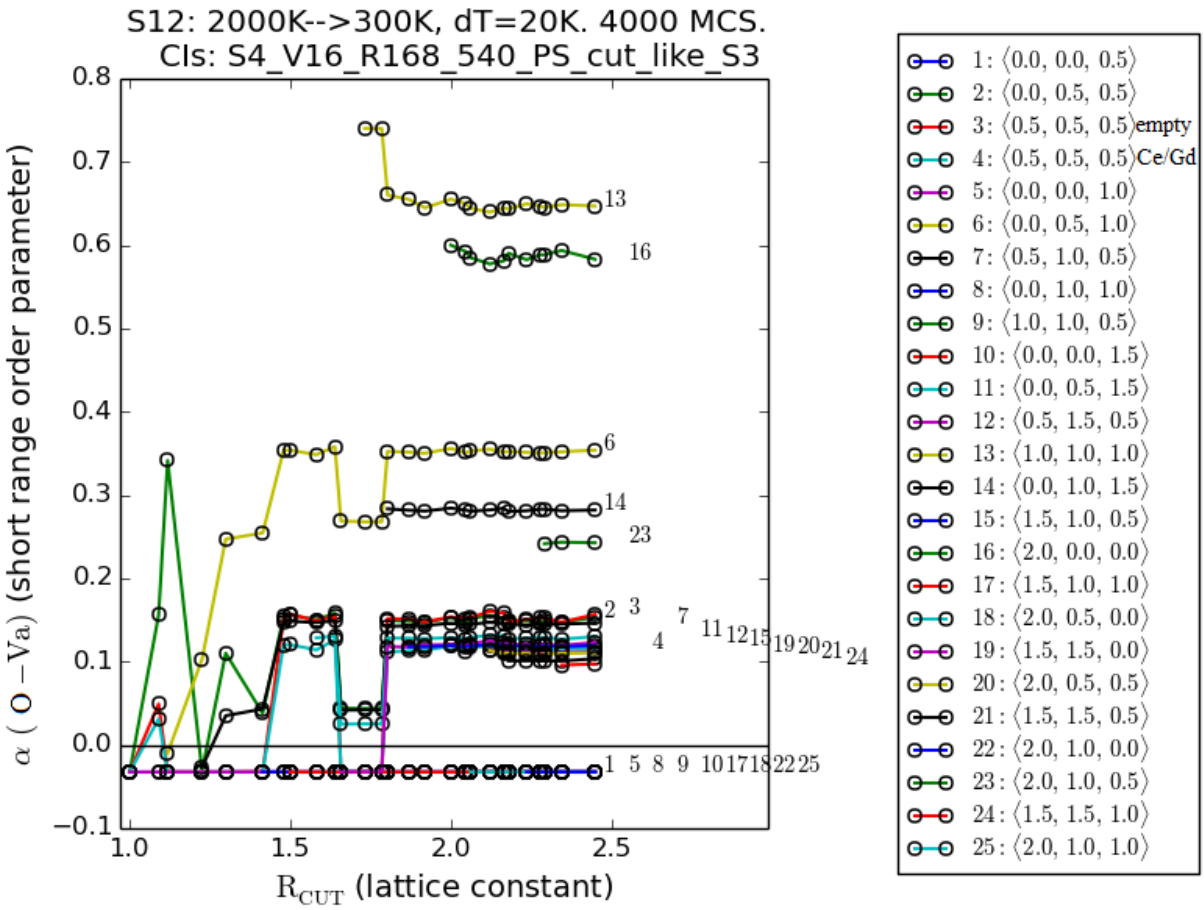

Figure S20: OVa SRO parameters (at 300 K) *vs.* cut-off radius of MC simulation ($x_{\mathrm{Gd}} = 0.125$, $12 \times 12 \times 12$ supercell, annealing from 2000 K to 300 K, 20 K temperature step, 4000 Monte Carlo steps per temperature). CIs obtained using cut-off radius like in the S3 case (thus CIs with $R > R^{\mathrm{CE}}_{\mathrm{CUT}}$ are zeros in MC). Legends show coordination shells (O&Va sublattice). Labels of vectors (shells) are written on the same levels (y coordinates) as the most long-ranged values of α_i's.

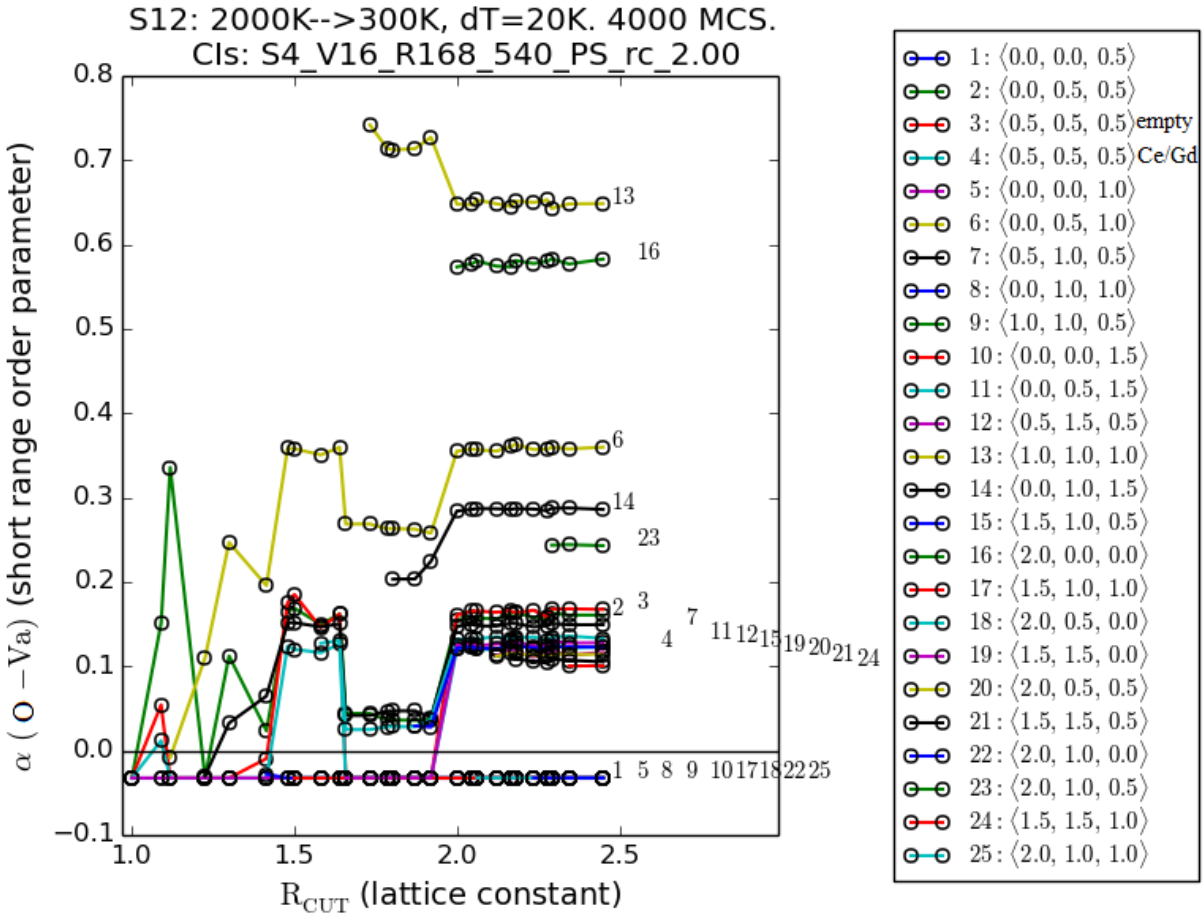

Figure S21: OVa short range order parameters (at 300 K) *vs.* cut-off radius of MC simulation ($x_{\mathrm{Gd}} = 0.125$, $12 \times 12 \times 12$ supercell, annealing from 2000 K to 300 K, 20 K temperature step, 4000 Monte Carlo steps per temperature). CIs obtained using cut-off radius 2.0 (thus CIs with $R > R^{\mathrm{CE}}_{\mathrm{CUT}}$ are zeros in MC). Legends show coordination shells (O&Va sublattice). Labels of vectors (shells) are written on the same levels (y coordinates) as the most long-ranged values of α's.

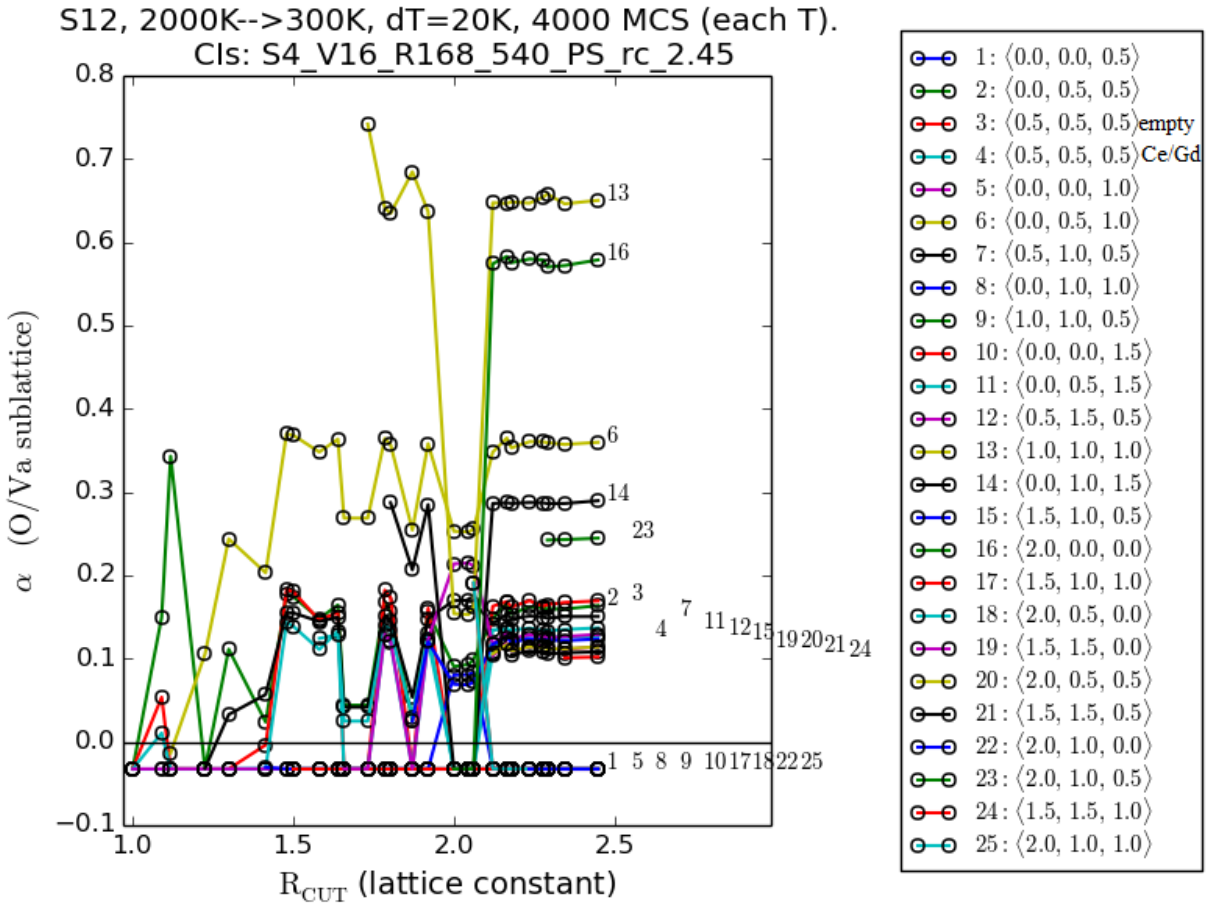

Figure S22: OVa short range order parameters (at 300 K) *vs.* cut-off radius of MC simulation ($x_{\mathrm{Gd}} = 0.125$, $12 \times 12 \times 12$ supercell, annealing from 2000 K to 300 K, 20 K temperature step, 4000 Monte Carlo steps per temperature). CIs obtained using cut-off radius $2.45a$. Legends show coordination shells (O&Va sublattice). Labels of vectors (shells) are written on the same levels (y coordinates) as the most long-ranged values of α's.

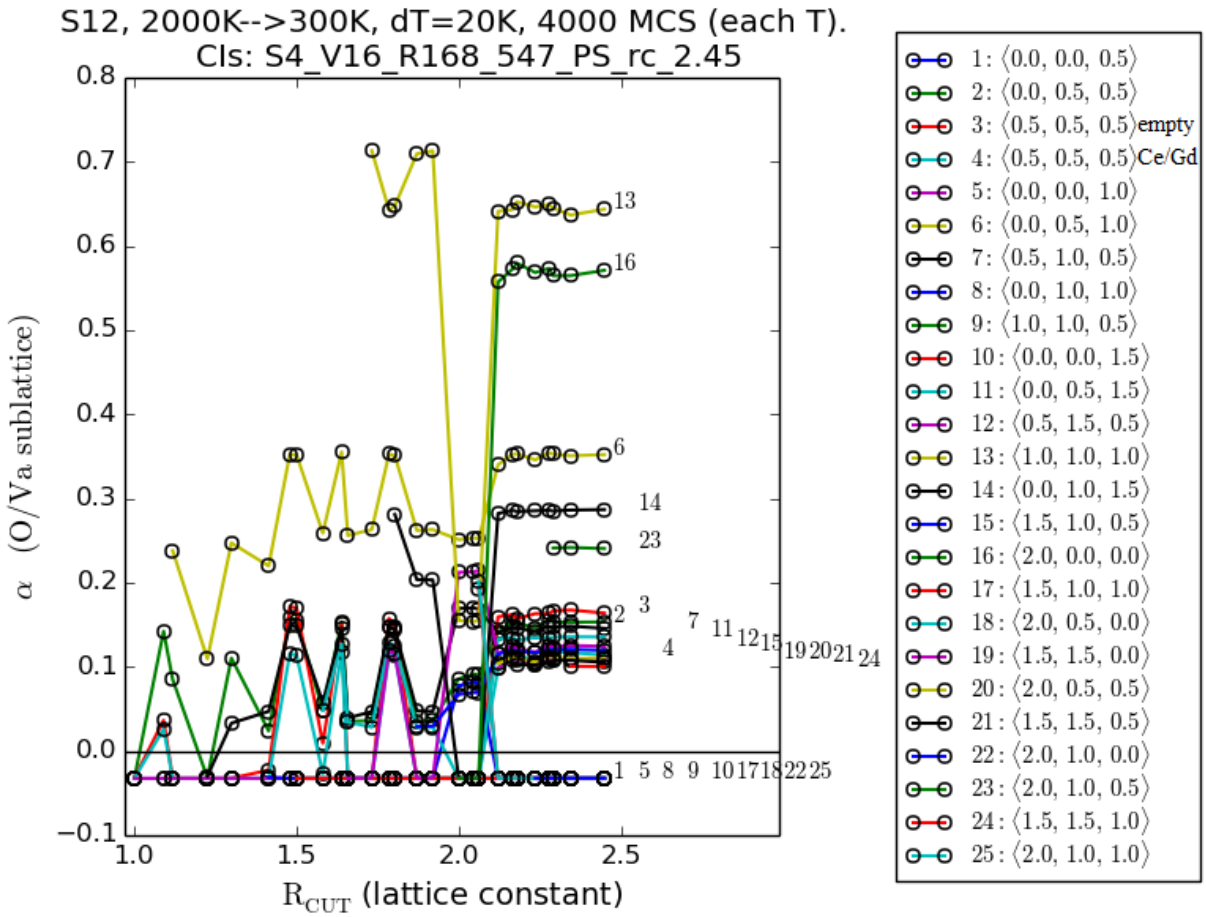

Figure S23: OVa SRO parameters (at 300 K) *vs.* cut-off radius of MC simulation ($x_{\mathrm{Gd}} = 0.125$, $12 \times 12 \times 12$ supercell, annealing from 2000 K to 300 K, 20 K temperature step, 4000 Monte Carlo steps per temperature). CIs obtained using cut-off radius $2.45a$. Notice that $a = 5.470$ Å. Legends show coordination shells (O&Va sublattice). Labels of vectors (shells) are written on the same levels (y coordinates) as the most long-ranged values of α's.

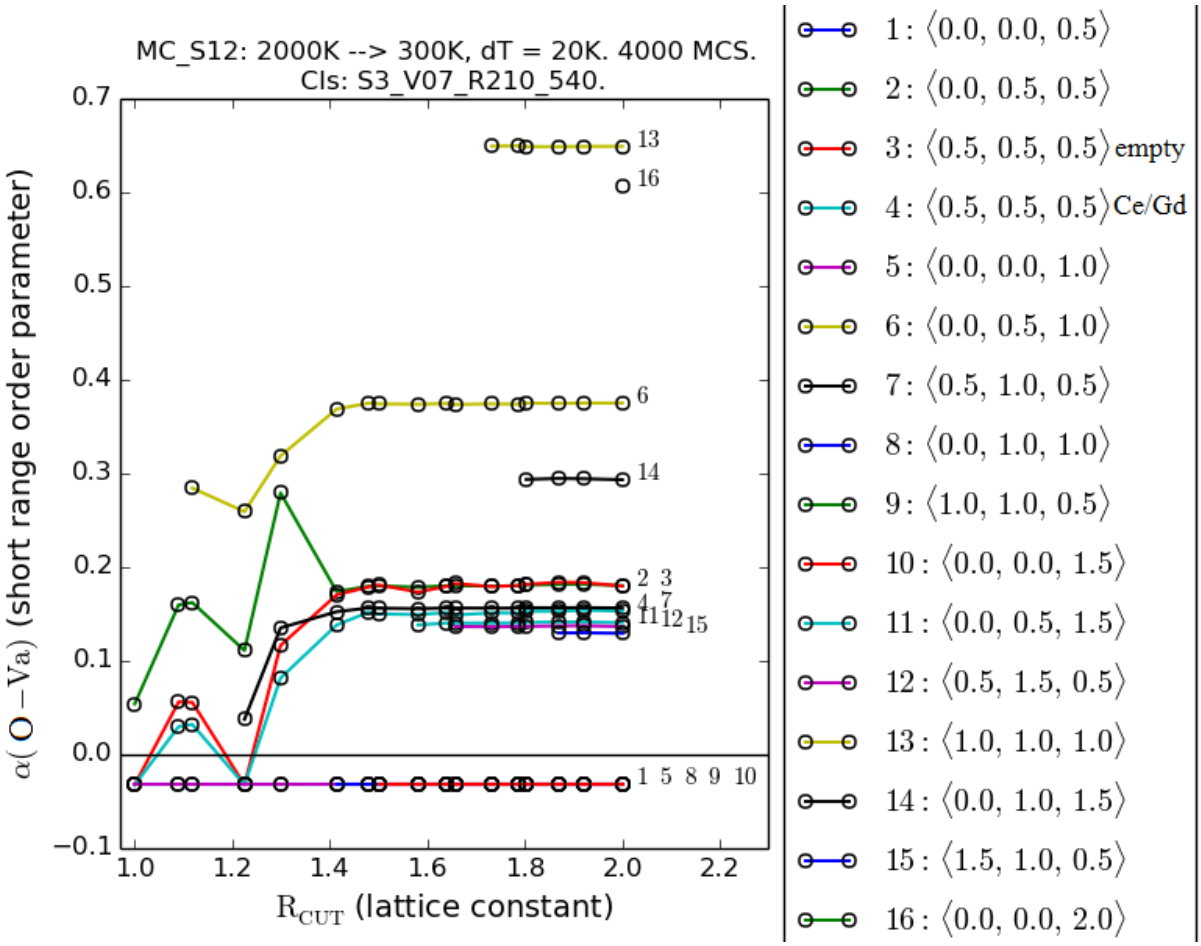

Figure S24: OVa short range order parameters (at 300 K) *vs.* cut-off radius of MC simulation ($x_{\mathrm{Gd}} = 0.125$, $12 \times 12 \times 12$ supercell, annealing from 2000 K to 300 K, 20 K temperature step, 4000 Monte Carlo steps per temperature). CIs obtained using cut-off radius usual for S3 (using set S3_07_R210_PS_a_5.40). Legends show coordination shells (O&Va sublattice). Labels of vectors (shells) are written on the same levels (y coordinates) as the most long-ranged values of α's.

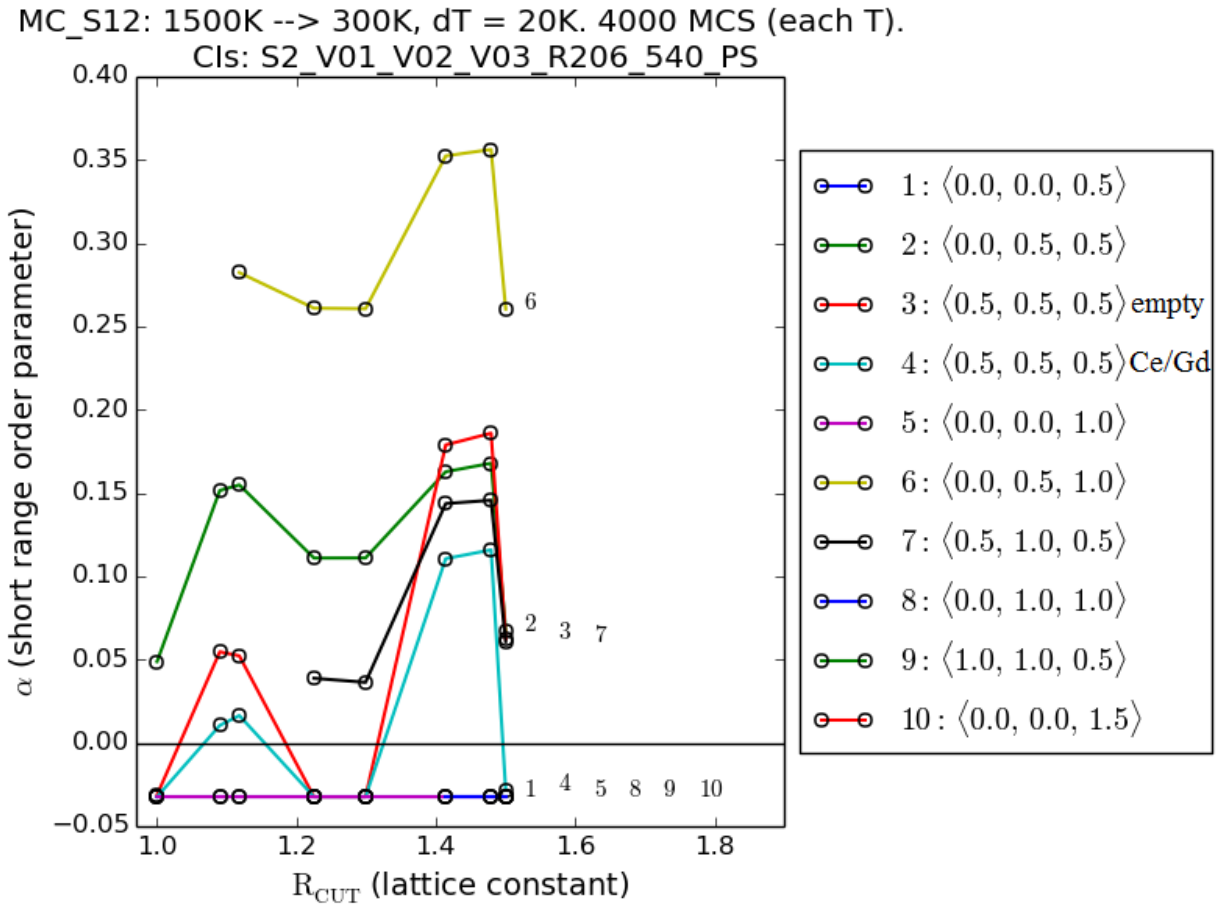

Figure S25: OVa SRO parameters (at 300 K) *vs.* cut-off radius of MC simulation ($x_{\mathrm{Gd}} = 0.125$, $12 \times 12 \times 12$ supercell, annealing from 1500 K to 300 K, 20 K temperature step, 4000 Monte Carlo steps per temperature). CIs obtained using cut-off radius usual for S2 (using set S2_V01-V02-V03_R206_PS_a_5.40). Legends show coordination shells (O&Va sublattice). Labels of vectors (shells) are written on the same levels (y coordinates) as the most long-ranged values of α's.

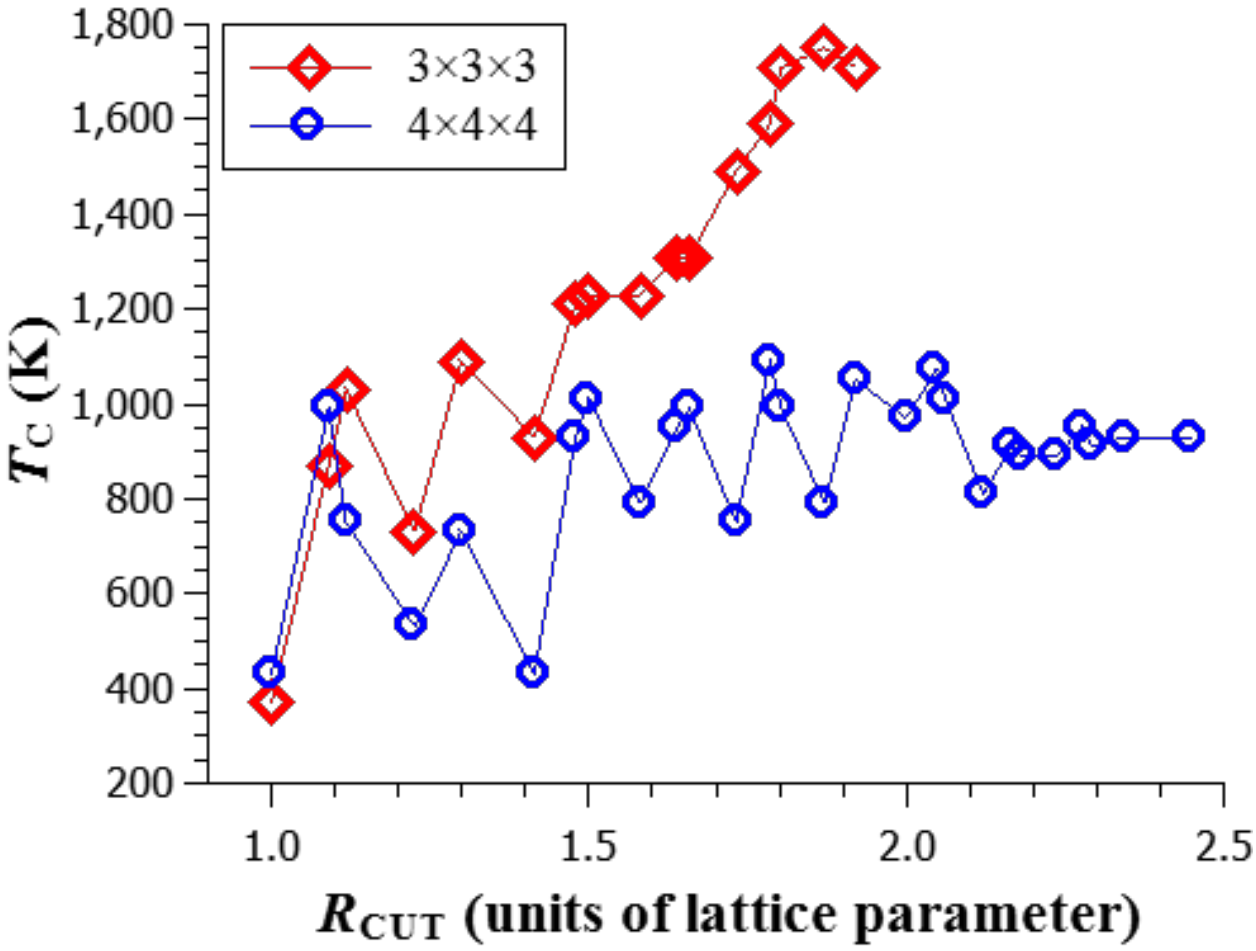

Figure S26: T_C *vs.* R_{CUT}^{MC} for S3_V07_R210_PS_a_5.47 and S4_V16_R168_PS_a_5.47 (denoted as $3 \times 3 \times 3$ and $4 \times 4 \times 4$, respectively).

2.b Reproducing C-type ordering in Gd_2O_3

The structure of C-type Gd_2O_3 and, therefore, Va−Va clustering is known from experiments, providing a test for our CE+MC modeling. The C-type Gd_2O_3 belongs to the *Ia-3* space group, with the following Wyckoff positions:[5] Gd no. 1: $24d$ $(x,\ 0,\ 1/4)$ $[x = 0.2815]$, Gd no. 2: $8a$ $(0,\ 0,\ 0)$, O no. 1: $48e$ $(x,\ y,\ z)$ $[x = 0.0985,\ y = 0.3628,\ z = 0.1287$, full occupancy], O no. 2: $16c$ $(x,\ x,\ x)$ $[x = 0.125$, zero occupancy]. Thus the C-type Gd_2O_3 unit cell can be seen as the ideal $2 \times 2 \times 2$ CeO_2 supercell with displaced Gd no. 1 and O no. 1, and missing O no. 2, therefore, it can be approximated with the same lattice model and Hamiltonian.

In short, our CE+MC approach allows to reproduce the correct C-type ordering in Gd_2O_3. The MC simulations of Gd_2O_3 (see Fig. S27) (temperature was gradually reduced from 4000 K to 500 K with a 50 K step) showed the following: 1) below ≈ 3000 K the ordered Gd_2O_3 phase formed (slight disorder due to thermal fluctuations was detected); 2) below ≈ 1500 K all vacancies were perfectly aligned: infinite non-crossing Va−Va chains in the $\langle 111 \rangle$ directions were formed (*i.e.* O no. 2 is missing). This clustering can also be seen as a zigzag-like network of vacancies joint with $\left\langle 1, \frac{1}{2}, 0 \right\rangle$ vectors.

In the case when the CIs obtained from the $2 \times 2 \times 2$ supercells were used, C-type ordering in Gd_2O_3 was exactly reproduced only with $R^{\mathrm{MC}}_{\mathrm{CUT}} = 1.479a, 1.5a$ (see Figs. S28). With smaller cut-off radii the ordering is not complete.

In the case when the CIs obtained from the $4 \times 4 \times 4$ supercells were used, the convergent solution was obtained (see Fig. S27).

Thus, these tests show that the stabilisation of the C-type ordering in Gd_2O_3 phase occurs only with $R^{\mathrm{MC}}_{\mathrm{CUT}} \gtrsim 1.5a$ (see Figs. S27, S28). This supports the conclusions of Sections 1.h and 2.a. In principle, one can obtain such a result using $2 \times 2 \times 2$ supercells, but, since for the $x_{\mathrm{Gd}} = 0.125$ case convergence is bad, we conclude that results based on DFT calculations with $2 \times 2 \times 2$ are not trustworthy.

[5] See C. Artini, M. Pani, A. Lausi, R. Masini, and G. A. Costa, "High Temperature Structural Study of Gd-Doped Ceria by Synchrotron Xray Diffraction (673K $\leq T \leq$ 1073K)", *Inorg. Chem.*, **53**, 10140-10149, 2014.

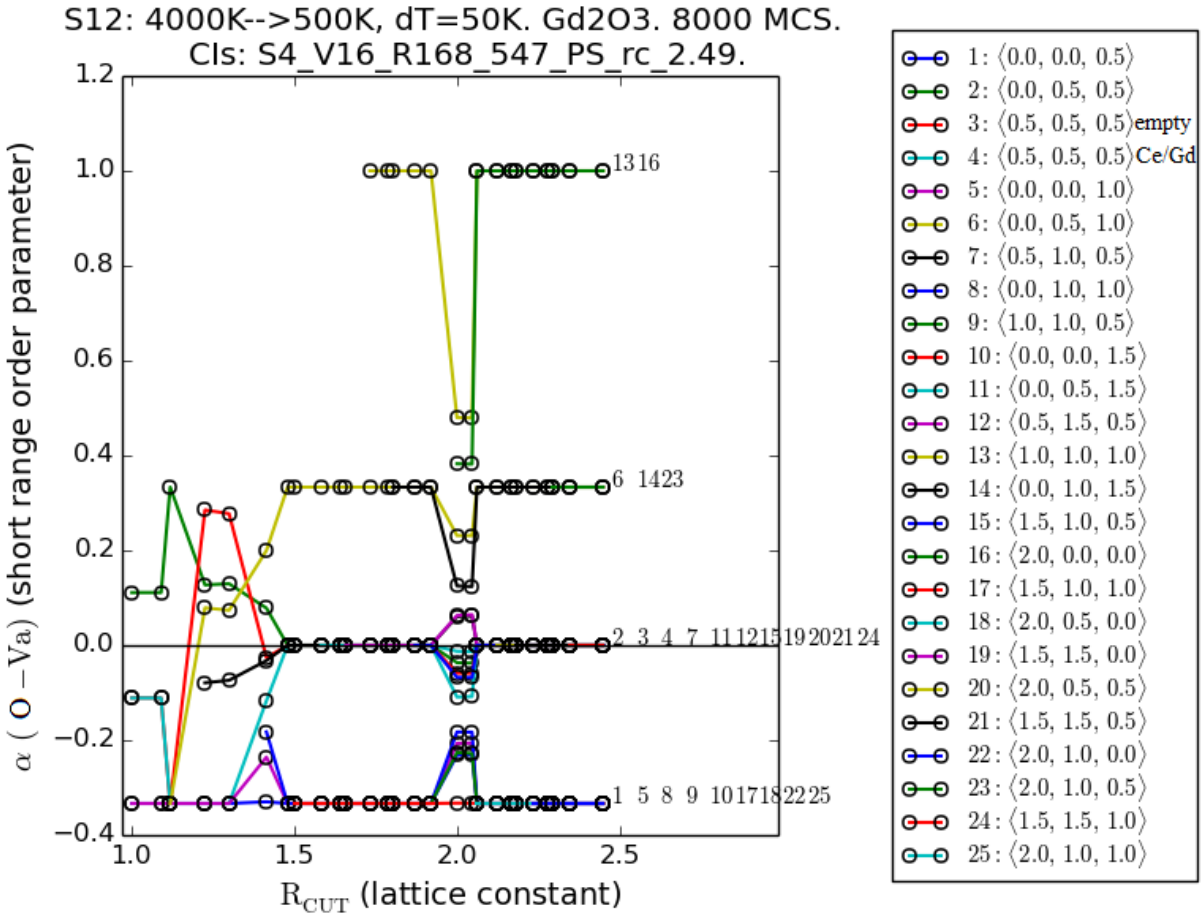

Figure S27: OVa short range order parameters (at 500 K) *vs.* cut-off radius of MC simulation of Gd_2O_3 ($x_{Gd} = 1$, $12\times12\times12$ supercell, annealing from 4000 K to 500 K, 50 K temperature step, 8000 Monte Carlo steps per temperature). CIs obtained using cut-off radius $2.45a$ (using set S4_V16_R168_PS_a_5.40). Legends show coordination shells (O&Va sublattice). Labels of vectors (shells) are written on the same levels (y coordinates) as the most long-ranged values of α's.

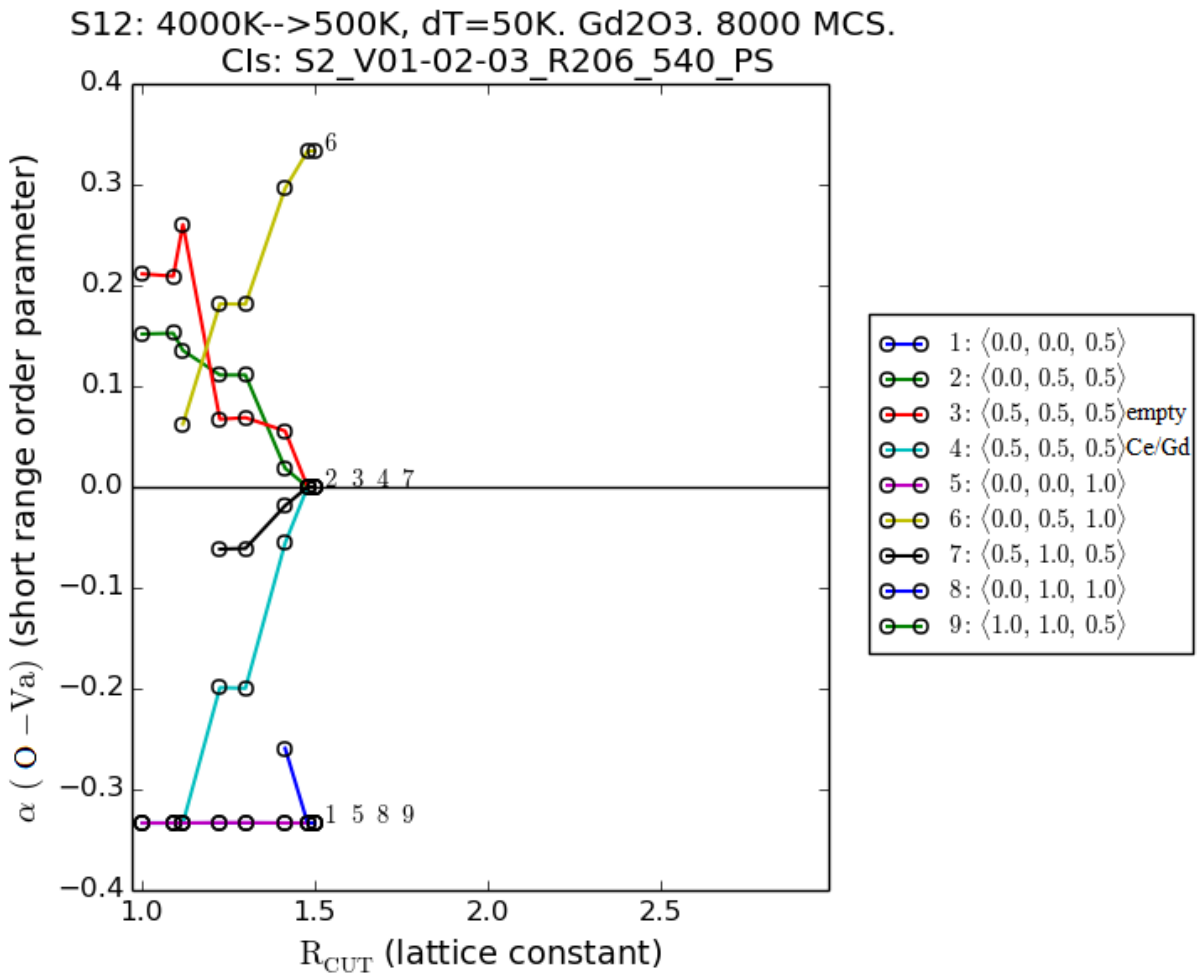

Figure S28: OVa short range order parameters (at 500 K) *vs.* cut-off radius of MC simulation of Gd_2O_3 ($x_{Gd} = 1$, $12\times12\times12$ supercell, annealing from 4000 K to 500 K, 50 K temperature step, 8000 Monte Carlo steps per temperature). CIs obtained using cut-off radius $1.5a$ (using set S2_V01-02-03_R206_PS_a_5.40). Legends show coordination shells (O&Va sublattice). Labels of vectors (shells) are written on the same levels (y coordinates) as the most long-ranged values of α's.

2.c MC tests regarding number of steps

Here we use the CIs obtained from set S4_V16_R168_a_5.47_PS using $R_{\mathrm{CUT}}^{\mathrm{CE}} = 2.45a$ and $R_{\mathrm{CUT}}^{\mathrm{MC}} = 2.1749a$. Figs. S29, S30, S31 show the energy evolution during MC simulated annealing.

- Starting from a random configuration well above the transition temperature (1500 K) the equilibration is reached after ≈ 500 MC steps (MCS) (see Fig. S29).
- At the the T_{C}, order-disorder transition occurs during $\gtrsim 3000$ steps. Just above the transition temperature ($\approx T_{\mathrm{C}} + 20$ K) large energy fluctuations might occur. (See Fig. S30.)
- Starting from a random configuration well below the transition temperature (720 K) the decomposition occurs in ≈ 2000 MCS (Fig. S31). Starting from a structure ordered at 720 K, the equilibration at 700 K occurs in $\approx 500 - 1000$ MCS. (See Fig. S31.)

Thus, obviously, the longest equilibration time is required close to the transition temperature, where roughly 3000 MCS are required. Since the MC simulations with many long-range interactions are time consuming, we use 3000 MCS for equilibration as a reasonable compromise between the accuracy and computational efficiency. We also collect statistics for 5000 MCS after the equilibration and average the results obtained in 16 independent runs. The temperature step is 20 K. Since equilibration below and above the transition temperature happens much faster, we expect the maximal errors to occur around the transition temperature (± 20 K) and the error in the determination of the transition temperature, $\Delta T_{\mathrm{C}} \approx 20$ K. The same settings are applied in the simulations with fixed Ce/Gd lattice. We notice that a small temperature step and 8000 MC steps (in total per temperature) ensure the correct structure evolution and convergence. Correct structure evolution is also ensured in our above-mentioned test calculations with 4000 MC steps (per temperature) — in those we do not focus on the average properties, but rather the final state at 300 K (therefore we do not need 8000 MC steps per temperature).

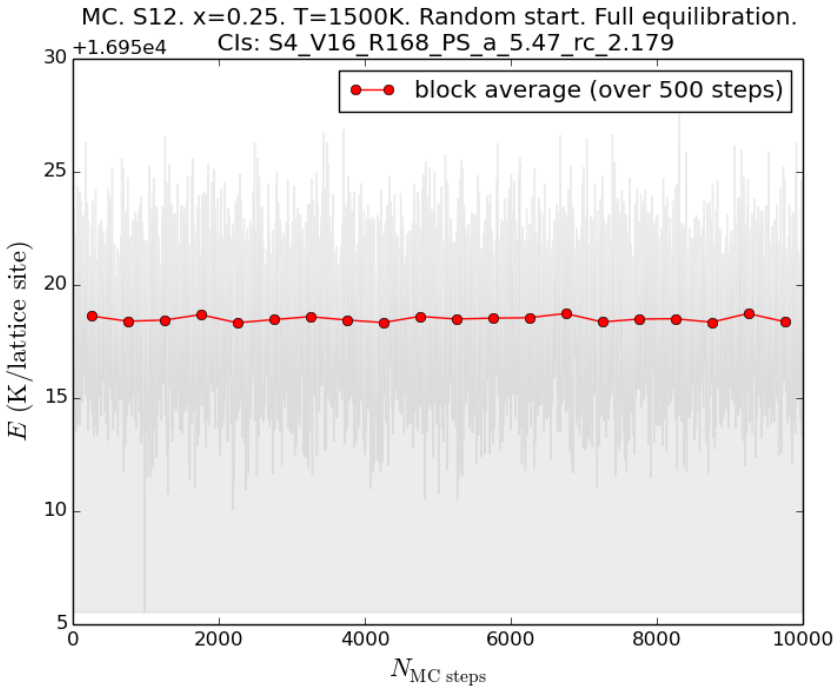

Figure S29: E change during MC run above the transition temperature. Block averaged energy (over 500 steps) is also shown.

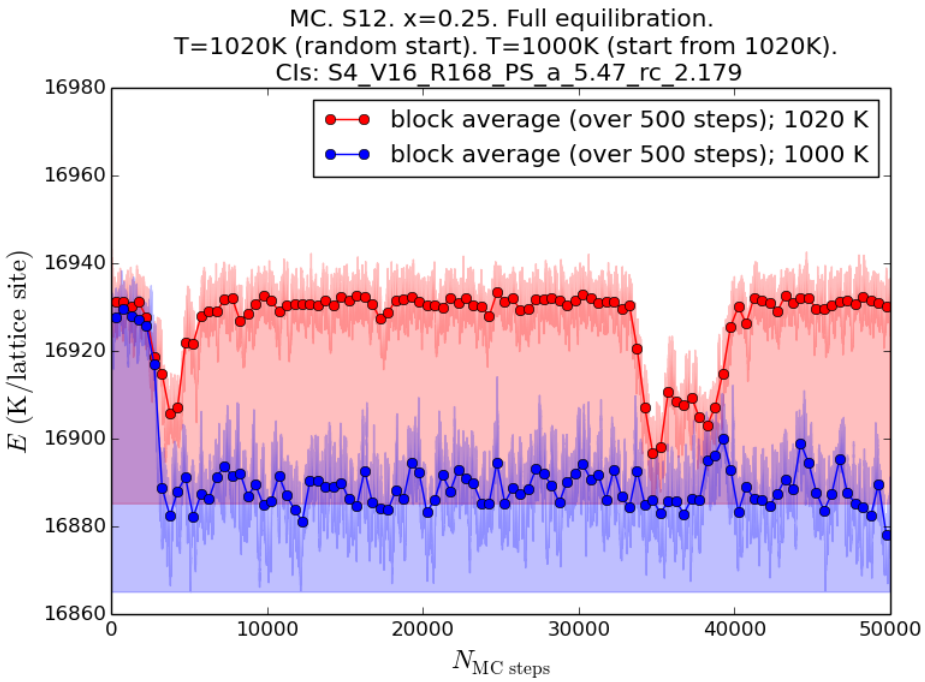

Figure S30: E change during MC run close to the transition temperature. Block averaged energy (over 500 steps) is also shown.

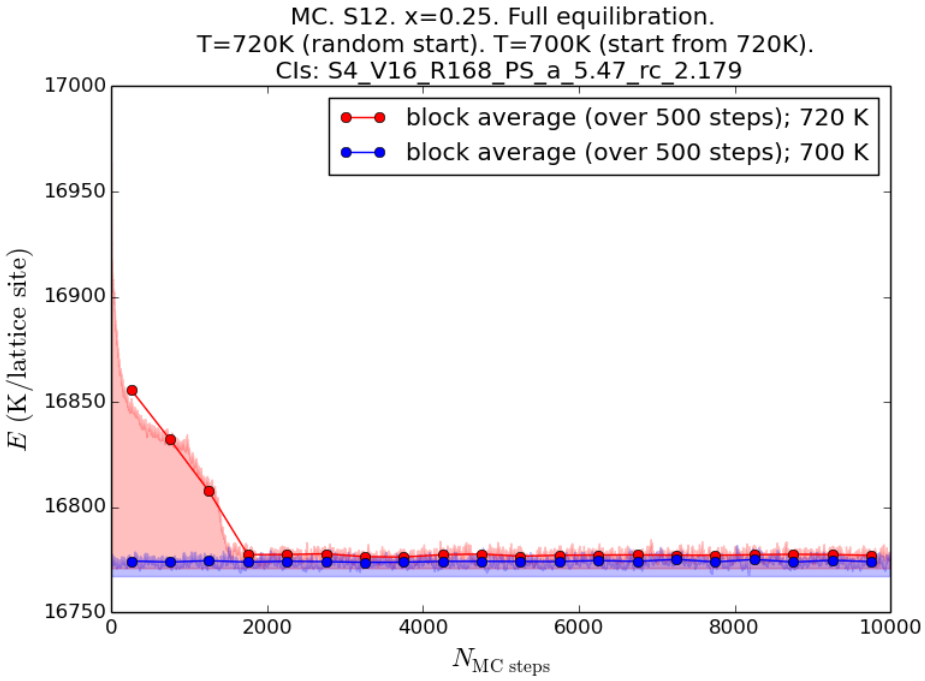

Figure S31: E change during MC run below the transition temperature. Block averaged energy (over 500 steps) is also shown.

2.d Convergence of T_C *vs.* supercell size in Monte Carlo simulations

First, let us notice that the correct ground state structure, *i.e.* decomposition into CeO_2 and C-type Gd_2O_3 below T_C, can be obtained in MC simulations using $6 \times 6 \times 6$ supercell or larger. For smaller supercells finite size comes into play and affects the ground state solution.

Regarding the form of precipitate it should be mentioned that, for, *e.g.* $x_{Gd} = 0.1875$, the spherical precipitate is formed using the $16 \times 16 \times 16$ supercell, and cylindrical in the case of the $12 \times 12 \times 12$ supercell. The form of precipitate does not affect the structure and is a minor issue.

Here we consider the convergence of T_C *vs.* supercell size N ($N \times N \times N$ supercell) for $x_{Gd} = 0.0625$ and 0.25 (see Fig. S32). From the extrapolations shown in Fig. S32 one can estimate the underestimation of the transition temperature. Thus for the $12 \times 12 \times 12$ supercell T_C is underestimated by ≈ 100 K ($x_{Gd} = 0.0625$) and ≈ 80 K ($x_{Gd} = 0.25$). For the $16 \times 16 \times 16$ supercell T_C is underestimated by ≈ 50 K ($x_{Gd} = 0.0625$) and ≈ 40 K ($x_{Gd} = 0.25$).

Since underestimation can be evaluated, we can use the $12 \times 12 \times 12$ supercells, which allows us to save computational resources.

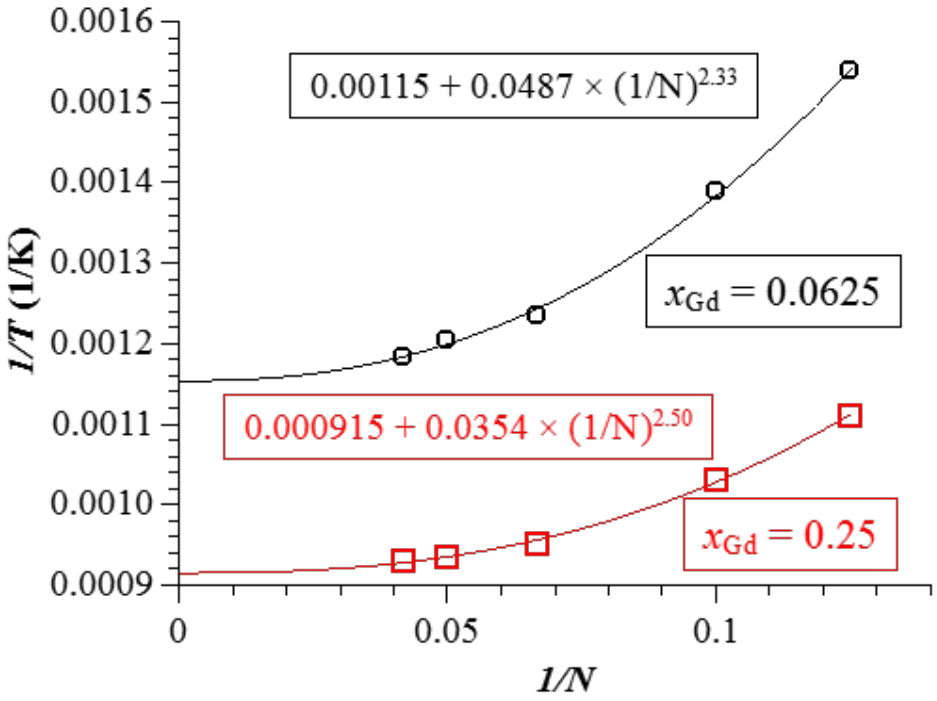

Figure S32: $1/T$ *vs.* $1/N$ for two concentrations. Extrapolations to $1/N \to 0$ ($N \to \infty$) are shown.

2.e Regarding ultimate Va trapping in the Gd_2O_3 precipitate

Fig. S33 shows that when the distribution of Gd atoms is fixed at the position obtained at 800 K (Gd_2O_3 precipitate exists), the oxygen vacancies are *ultimately* trapped. Above $\approx$ 2500 K O−Va ordering is more random-like. Notice, that even at 4000 K vacancies are strongly bound to Gd.

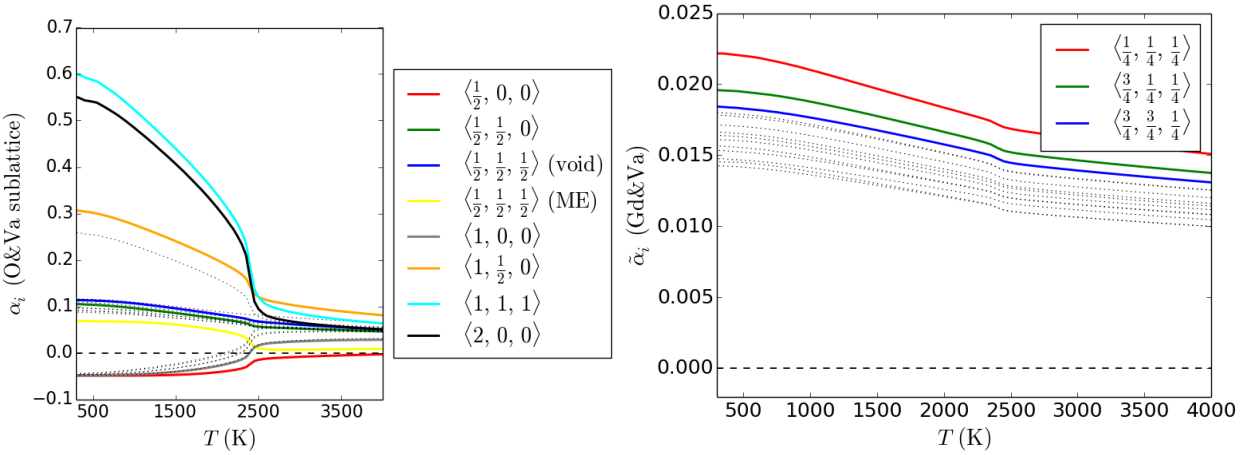

Figure S33: Fixed–Gd case ($T_{\mathrm{Gd}} = 800$ K), $x_{\mathrm{Gd}} = 0.1875$. Selected OVa and GdVa SRO parameters *vs.* T are shown in colour, while the rest with dotted lines. Dashed lines correspond to random distributions.