GTT-Technologies
Kaiserstraße 103
52134 Herzogenrath, Germany
Phone: +49-(0)2407-59533
Fax: +49-(0)2407-59661
E-mail: info@gtt-technologies.de

The *ab initio* Materials Project (AIMP) and OQMD (AIOQ) databases Version 5.0 Documentation

GTT-TECHNOLOGIES

Table of contents

Introduction

AIMP and AIOQ are developed by GTT-Technologies, based on provided *ab initio* data from Materials Project (materialsproject.org) [1-3] and Open Quantum Materials Database (OQMD.org) [4, 5] as well as own data and developed machine learning (ML) models as input.

The **AIMP** compound database v5.0 contains data for 128.978 solid phases that were calculated using *ab initio* methods in the context of the Materials Project (https://next-gen.materialsproject.org) [1-3] as of 3 February 2022. Derived from these 128.978 solid phases, a total of 88.373 compounds are introduced, with numerous compounds having multiple calculated crystallographic structures that are introduced as separate phases into the database. 44.001 phases contained in AIMP are experimentally observed and take ICSD (Inorganic Crystal Structure Database) entries as input. Besides stoichiometric compound data, the AIMP solutions database v5.0 contains data for metallic FCC, BCC, HCP solid solutions. Mixing enthalpies of these solid solutions have been obtained from *ab initio* calculations by GTT-Technologies.

The **AIOQ** database contains results from ground state *ab initio* calculations calculated by Chris Wolverton's group at Northwestern University (www.oqmd.org) [4, 5]. As of 3 February 2022, the latest Open Quantum Materials Database (OQMD) version was v1.5. The AIOQ database contains data for 718.923 solid phases and a total of 388.118 compounds. 32.079 of those phases in AIOQ are based on experiments and take ICSD entries as input.

Both repositories (Materials project and OQMD) only contain results from *ab initio* calculations at 0 K. The models developed by GTT and used to estimate thermodynamic properties at temperatures above 298 K are described in the later chapters after remarks on validity and possible application areas.

Validity and Applicability

Unlike all other databases available in FactSage, AIMP and AIOQ contain non-curated data. Therefore, it cannot be expected that *ab initio* databases yield the same accuracy as other curated FactSage databases. However, they extend the otherwise very limited composition space covered by curated databases opening up new possibilities for exploration. There are a few limitations: AIMP and AIOQ do not contain data for liquid; Gibbs energies are modelled only above room temperature; magnetic phases can have incorrect Gibbs energies since magnetism is not covered.

Using data analytics, all formation enthalpies, entropies as well as heat capacities have been checked to be generally reasonable and acceptable given the inaccuracies of the first principles methods, machine learning models and experimental input data that were used. As mentioned below, most formation enthalpies have been corrected based on data in existing FactSage databases.

There are several major applications for these databases:

- Use AIMP/AIOQ data as a starting point for a CALPHAD assessment in the Calphad Optimizer (available since FactSage 8.2) to complement incomplete experimental data.
- Combine standard FactSage databases with AIMP and AIOQ to estimate thermochemical properties in parts of chemical compound space where otherwise no data is available to describe the behavior of minor elements.
- Scan systematically complete chemical space using only AIMP/AIOQ in a materials informatics approach to identify interesting materials.

Applied models

In the following, the machine learning (ML) models to estimate the thermodynamic properties at temperatures above 298 K of all phases contained in AIMP as well as AIOQ are described. Individual models for heat capacity, entropy and formation enthalpy modifications are developed at GTT which are introduced in the given order. To the best of our knowledge, there are no comparable models trained on equally large training sets.

Heat Capacity

Heat capacity C_p is estimated with several Gaussian process regressions that are applied at selected finite temperatures. The models are trained with thousands of pure compounds which exist in FactSage databases while simultaneously having an entry in Materials Project. Phases metastable at room temperature are avoided, i.e., only stable phases under standard conditions are used for the training. At each temperature, a different Gaussian process is trained. C_p is thus estimated at 10 different temperatures. On top of these estimated values, the heat capacity function

$$C_p(T) = A + BT + CT^2 + DT^{-2}$$

is fitted using Levenberg–Marquardt algorithm. The temperature range of the function is defined between 298 K and 5000 K for all phases. Provisions are in place to prevent unrealistic extrapolations at very high temperatures (above liquidus temperature).

In the following Figure 1, examples for three representative compounds (an intermetallic, a carbide and an oxide) are provided. On the left, the heat capacity obtained from the 10 different Gaussian processes are shown together with the fit. On the right, a comparison to data existing in different FactSage databases is shown. Please note that these compounds were not considered in training of the models by any means.

In all three cases, excellent agreement is observed between ML predicted and fitted heat capacities while the fitted heat capacities show similar agreement to the different FactSage databases as these agree with each other. In the case of $NaAlSi_3O_8$, the unreasonable high heat capacity assumed in AIMP4.0 is avoided in this version.

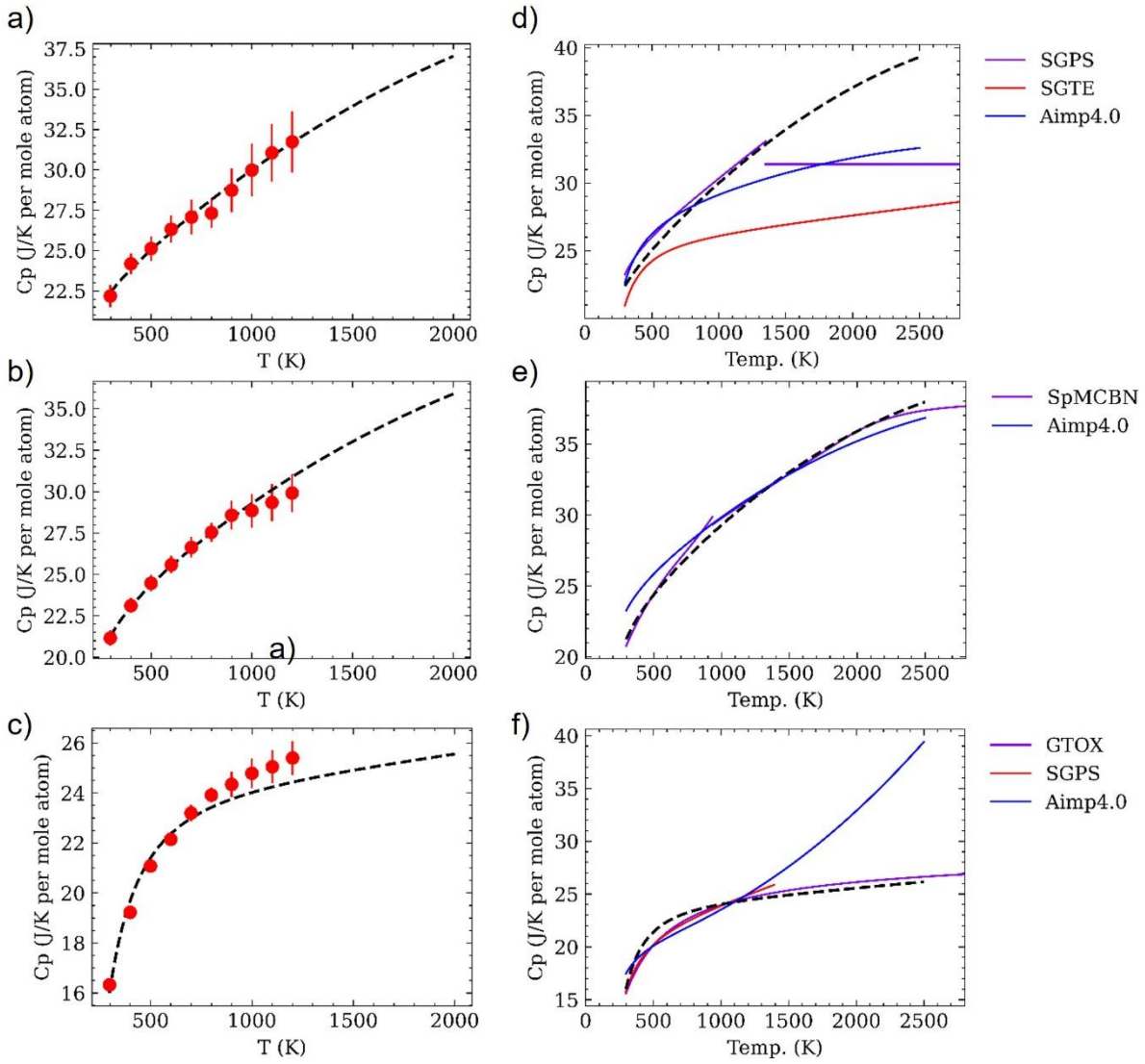

Figure 1: Heat capacity of Mg_2Si (a, d), Ti_2AlC (b,e) and $NaAlSi_3O_8$ (c,f). (a-c): Red dots are the results of the GP regressions. Error bars are the uncertainties. Black dashed line is the corresponding fitted heat capacity function in AIMP v5.0. (d-f): Dashed line is the fitted heat capacity function in AIMP v5.0, same as in a-c. Other functions are from different FactSage databases.

Entropy

Similar to heat capacity, entropy S is estimated with several Gaussian process regressions that are applied at selected finite temperatures. Finally, entropy is described using the C_p model described above as well as S^{298}. Comparing S^{298} predictions here with experimental data, the calculated mean absolute error (MAE) of a test set containing 480 randomly selected compounds is 2.1 J/K per mole atom and root mean squared error (RMSE) is 3.1 J/K per mole atom. In comparison, uncertainty

of experimentally determined S298 is around 1.1 J/K per mole atom according to the data provided by Kubaschewski [6].

A comparison for the same three compounds -as shown in Figure 1 with respect to heat capacity- between ML predictions of entropy at different temperatures, fitted data and data in existing FactSage databases is shown in Figure 2. Again, agreement as good as the one between different databases in FactSage is observed.

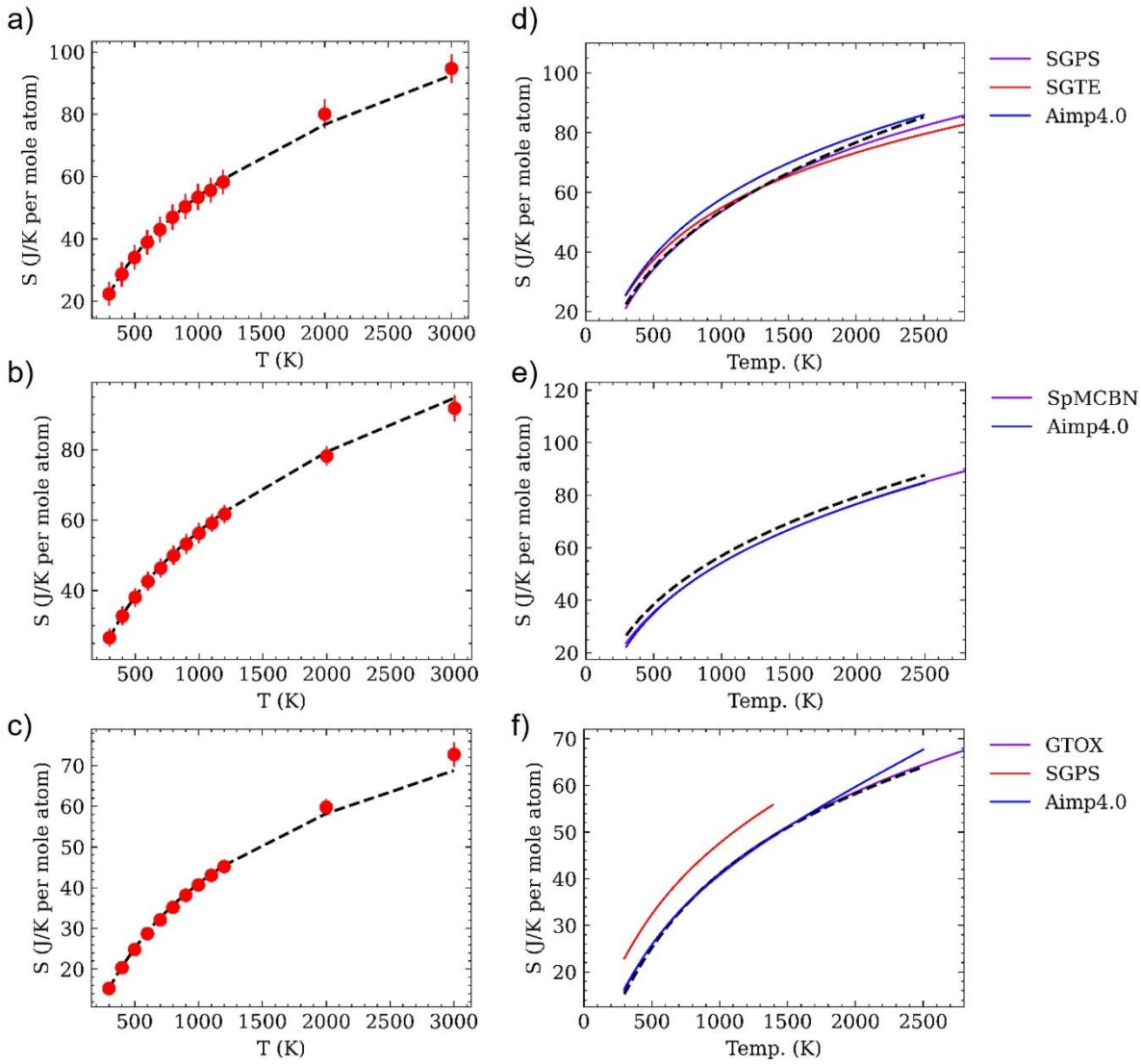

Figure 2. Entropy of Mg_2Si (a, d), Ti_2AlC (b,e) and $NaAlSi_3O_8$ (c,f). (a-c): Red dots are the results of the GP regressions. Error bars are the uncertainties. Black dashed line is the corresponding fitted entropy function in AIMP v5.0. (d-f): Dashed line is the fitted entropy function in AIMP v5.0, same as in a-c. Other functions are from different FactSage databases.

Formation enthalpy modifications

Approximate DFT functionals result in systematic errors and a correction of DFT formation enthalpies is required. Even though MP and OQMD are already applying such corrections, systematic deviations to experimental data are still observed. This is partly due to the insufficiently large training datasets used by MP and OQMD. Thus, we apply additional modifications to the formation enthalpy values provided by materialsproject.org and oqmd.org. Figure 3 shows the Materials Project versus experimental formation enthalpies of compound sets of nitrates and sulfates, both of which show large deviations when comparing to experimental enthalpy of formation data.

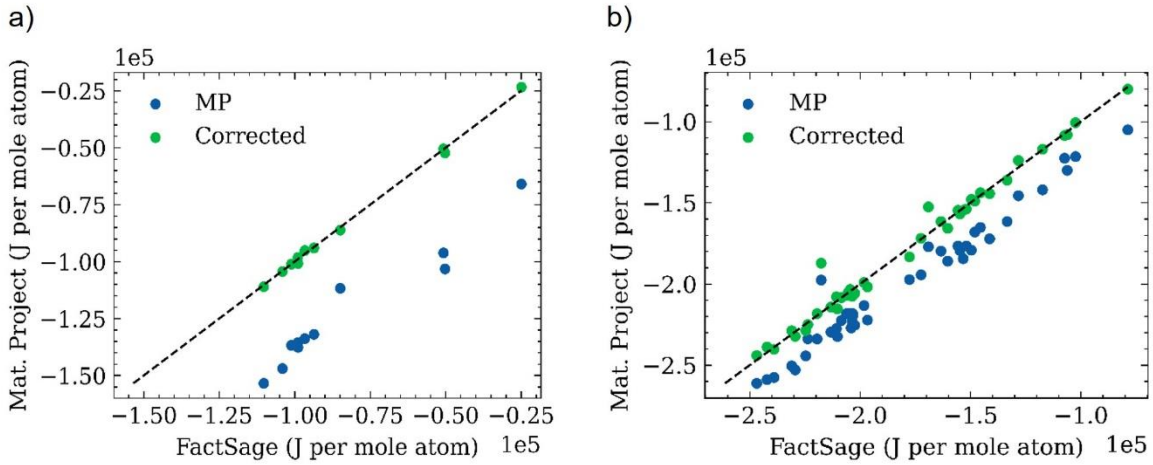

Figure 3. (a): Comparison between Materials Project and FactSage formation enthalpies of nitrates. There exists a systemic error even though enthalpies provided by Materials Project already have a correction scheme. (b): Similar to (a), systemic error is also observed between formation enthalpies of sulphates.

We are following a similar approach to Materials Project correction scheme [7]. However, instead of considering a few hundred compounds, our training set contains 3547 compounds. The mean absolute error between calculated and experimental formation enthalpies is around 15 kJ per mole atom for both MP and OQMD datasets. We reduce this number to 10 kJ per mole atom by increasing the corrected number of ions in the model. Two different models, specified for Materials Project and OQMD datasets, are trained.

It should be noted that 10 kJ per mole atom is still significantly higher than the desired accuracies. To give an example, shifting the formation enthalpy of a compound by only 1 kcal per mole atom (4.2 kJ per mole atom) while keeping the liquid free

energy curve the same leads to a ~500 K shift of congruent melting point. Thus, phase diagrams exhibiting the same accuracy as in case of using curated FactSage databases can in general not be expected.

Density and Elastic properties

For all compounds, the density is calculated from the relaxed unit cell volume given in Materials Project or OQMD and is included as constant in the databases. It should be noted that the Generalized Gradient Approximation (GGA) used in the *ab initio* calculations results in a systematic error, overestimating the unit cell volume, i.e., underestimating density [8].

For some phases the elastic constants can be extracted from Materials Project.

Included elements and discarded phases

There are 76 elements included in both Materials Project and OQMD compounds. These elements are:

Ag, Al, As, Au, B, Ba, Be, Bi, Br, C, Ca, Cd, Ce, Cl, Co, Cr, Cs, Cu, Dy, Er, Eu, F, Fe, Ga, Gd, Ge, H, Hf, Hg, Ho, I, In, K, La, Li, Lu, Mg, Mn, Mo, N, Na, Nb, Nd, Ni, O, P, Pb, Pd, Pr, Pt, Pu, Rb, Re, Rh, Ru, S, Sb, Sc, Se, Si, Sm, Sn, Sr, Ta, Tb, Te, Th, Ti, Tm, U, V, W, Y, Yb, Zn, Zr.

These elements are the most important ones of technical interest. Other elements are not included due to the lack of data to benchmark and unfeasible training of ML models.

Additionally,

- Phases tagged as amorphous are discarded.
- Phases exhibiting a volume larger than 120 $Å^3$ and smaller than 4 $Å^3$ per atom are discarded.
- If there are more than 50 phases exist for a compound, the 50 with lowest formation enthalpy are included.
- Compounds exhibiting more than 7 elements are discarded.

Solid solutions

Enthalpies of mixing at 0 K have been systematically calculated by GTT-Technologies for the FCC_A1, BCC_A2 and HCP_A3 solutions. Based on these, 1900 binary interaction parameters have been derived.

In the phase FCC_A1, 1146 interaction parameters have been derived for binary systems combining any metal with atomic number between 3 (Li) and 83 (Bi) with one of the following elements: Al, Ca, Ni, Cu, Sr, Rh, Pd, Ag, Ir, Pt, Au, Pb

In the phase BCC_A2, 608 interaction parameters have been derived for binary systems combining any metal with atomic number between 3 (Li) and 83 (Bi) with one of the following elements: Li, Na, K, V, Fe, Nb, Mo, Ta, W

In the phase HCP_A3, 146 interaction parameters have been derived for binary systems combining any metal with atomic number between 3 (Li) and 83 (Bi) with one of the following elements: Mg, Ti, Zr.

Files

AIMP and AIOQ are delivered together. They contain the following files which should reside in the FACTDATA folder of FactSage:

- AIMPbase.cdb = compound database of AIMP
- AIMPbase.xcl = exclusion list, an ASCII file taking care of de-selecting pure elements from AIMP when it is used together with ELEM or SGUN, which contain better descriptions of the elements.
- AIMPsoln.sdc and AIMPsoln.fdb = AIMP solution database containing data for FCC_A1, BCC_A2 and HCP_A3.
- OQMEbase.cdb = compound database of AIOQ, containing only intermetallic compounds
- OQNMbase.cdb = compound database of AIOQ, containing only non-metallic compounds
- OQMEbase.xcl and OQNMbase.xcl = exclusion lists as for AIMP.

What is new?

Materials Project and OQMD databases are constantly being updated. Naturally, these changes are also applied to AIMP v5.0. AIOQ is newly introduced, since it uses essentially the same ML models as AIMP, it is also versioned as v5.0.

Compared to v4.0, estimation of heat capacities and entropies have further been significantly improved in v5.0. This has been achieved by extending the model training sets, and by detailing the models even further. Furthermore, phonon calculations are excluded in v5.0 since they have not proven to be consistently more reliable than the ML models developed here due to imaginary frequencies, fixed volume, lack of anharmonic effects, missing electronic contribution etc. Furthermore, a much more comprehensive enthalpy correction model is applied in v5.0now taking also ions into account.

References

[1] S. P. Ong *et al.*, “The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles,” *Comput. Mater. Sci.*, vol. 97, pp. 209–215, 2015.

[2] G. Ceder and K. Persson, “The materials project: A materials genome approach,” *DOE Data Explorer, http://www. osti. gov/dataexplorer/biblio/1077798.[2016-08-28]*. 2010.

[3] S. P. Ong *et al.*, “Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis,” *Comput. Mater. Sci.*, vol. 68, pp. 314–319, 2013.

[4] S. Kirklin *et al.*, “The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies,” *Npj Comput. Mater.*, vol. 1, no. 1, p. 15010, Dec. 2015, doi: 10.1038/npjcompumats.2015.10.

[5] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, “Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD),” *JOM*, vol. 65, no. 11, pp. 1501–1509, Nov. 2013, doi: 10.1007/s11837-013-0755-4.

[6] O. Kubaschewski, C. B. Alcock, P. J. Spencer, and O. Kubaschewski, *Materials thermochemistry*, 6th ed., rev.Enl. Oxford ; New York: Pergamon Press, 1993.

[7] A. Wang *et al.*, “A framework for quantifying uncertainty in DFT energy corrections,” *Sci. Rep.*, vol. 11, no. 1, p. 15496, Dec. 2021, doi: 10.1038/s41598-021-94550-5.

[8] A. Jain *et al.*, “A high-throughput infrastructure for density functional theory calculations,” *Comput. Mater. Sci.*, vol. 50, no. 8, pp. 2295–2310, Jun. 2011, doi: 10.1016/j.commatsci.2011.02.023.

Contact

In case of questions or to provide feedback, please open a new ticket in our support center: https://support.gtt-technologies.de/ or contact us via

GTT-Technologies
Kaiserstraße 103
52134 Herzogenrath, Germany
Phone: +49-(0)2407-59533
Fax: +49-(0)2407-59661
E-mail: info@gtt-technologies.de