The ab initio Materials Project (aiMP) and OQMD (aiOQ) v6.3 databases

Introduction

The aiMP database v6.3 contains data for 154.718 solid phases that were calculated by various groups using *ab initio* methods in the context of the Materials Project (https://next-gen.materialsproject.org) [1-3] as of 1 November 2023. Derived from these 154.718 structures, a total of 136.854 compounds are introduced, with a range of compounds having multiple calculated crystallographic structures that are introduced as separate phases into the database. The Materials Project repository contains results from *ab initio* calculations at 0 K and 0 atm. The models used to estimate thermodynamic properties at temperatures above 298 K are described in the later chapters after remarks on validity and possible application areas.

Similar to aiMP, 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 4 September 2023, the latest Open Quantum Materials Database (OQMD) version was v1.5. The AIOQ database contains data for 475'887 compounds.

aiMP and aiOQ are developed by GTT-Technologies, using data from Materials Project and OQMD as well as own data as input. They are complemented by the aiMP solutions database containing data for metallic FCC, BCC, HCP solid solutions. Phase stability of these solid solutions has been calculated from *ab initio* calculations by GTT-Technologies. The models used are described below.

Database files

The aiMP and aiOQ databases are split into smaller database files to make setup and evaluation of Equilib calculations easier. The following table shows the different database files, their contents and the respective use cases.

Database file and **FactSage nickname**	Contains	Number of phases	Use cases
AIMPsoln.SDC and AIMPsoln.FDB **AIMP**	Solid solutions: FCC, BCC, HCP	3	Application calculations, materials informatics, database development
AIMPbase.CDB **AIMP**	Stable phases of aiMP	102'171	Application calculations, materials informatics, database development
AIMMbase.CDB **AIMM**	Metastable phases of aiMP	34'683	Database development
AIOQbase.CDB **AIOQ**	Stable phases of aiOQ	341'526	Application calculations, materials informatics, database development
AIOMbase.CDB **AIOM**	Metastable phases of aiOQ	134'361	Database development

Validity and Applicability

Unlike all other databases available in FactSage, aiMP and aiOQ contain non-curated data. Therefore, *ab initio* databases cannot be expected to lead to as accurate results as it is the case when using other FactSage databases.

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 that were used. As mentioned below, most formation enthalpies have been corrected based on data in existing FactSage databases.

There are three major applications for these databases:

- Using as a starting point for a CALPHAD assessment.
- Combining 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.
- Materials informatics screening of chemical space, especially in connection with ChemApp for Python.

Stable and metastable databases

As stated above, aiMP and aiOQ databases were split into two different databases: AIMP/AIOQ contain the stable phases and should be used for application calculations; AIMM/AIOM contain the metastable phases and should be used for thermodynamic database development or if a known metastable polymorph is of interest. The criteria for a phase to be in the stable database are:

- The phase with the lowest enthalpy of formation at 298 K
- All phases with lowest Gibbs between 300K and 5000 K for each unique composition
- All phases that have a “exp” tag in materialsproject.org or oqmd.org, i.e. that are considered experimentally confirmed

Heat Capacity

Heat capacity C_p is estimated with several Gaussian process regressions that are applied at some finite temperatures. The models are trained with pure compounds which exist in any FactSage databases and have an entry in Materials Project. Metastable phases 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 estimated at 10 different temperatures and entropy (S) is estimated at 12 different temperatures. On top of these estimated values, 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. To prevent unrealistic extrapolations at elevated temperatures, liquid entropies are estimated at 4000 K and 5000 K. These values are also considered while fitting the heat capacity function.

In order to provide a better summary, we present 3 compounds from our test set. These compounds take no part in training of the models by any means. Figure 1 summarizes how heat capacity of MAX phase Ti_2AlC is obtained.

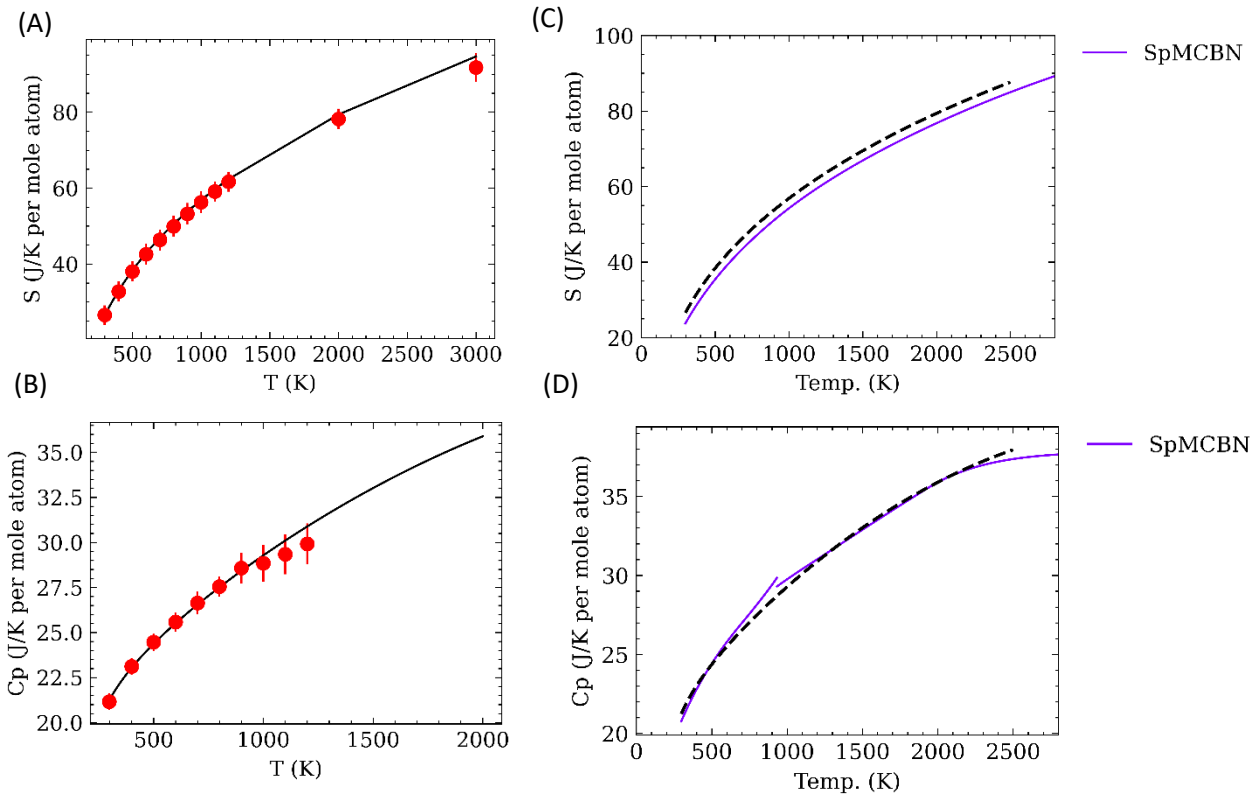

Figure 1. Heat capacity of Ti_2AlC. (A): Estimated entropies. Red dots are the results of the GP regressions. Error bars are the uncertainties. Black line is the corresponding entropy curve of the fitted heat capacity function. (B): Similar to (A), red dots and error bars are the result of GP regressions. Heat capacity function is fitted on both entropy and heat capacity points and is the black curve. (C): Dashed line is the estimated entropy curve, i.e., same black curve in (A). Other functions are from the FactSage databases. (D): Estimated heat capacity and heat capacities in FactSage databases.

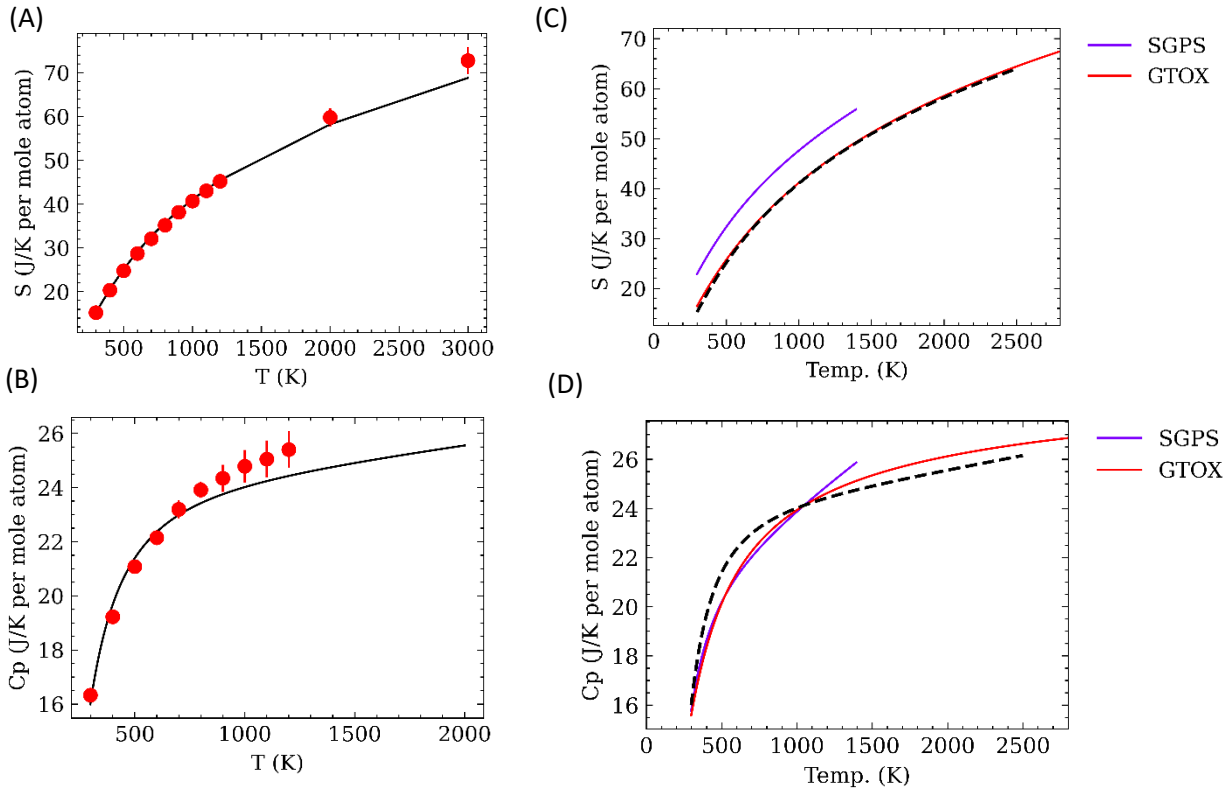

Figure 2. Heat capacity of $NaAlSi_3O_8$. (A): Estimated entropies. Red dots are the results of the GP regressions. Error bars are the uncertainties. Black line is the corresponding entropy curve of the fitted heat capacity function. (B): Similar to (A), red dots and error bars are the result of GP regressions. Heat capacity function is fitted on both entropy and heat capacity points and is the black curve. (C): Dashed line is the estimated entropy curve, i.e., same black curve in (A). Other functions are from the FactSage databases. (D): Estimated heat capacity and heat capacities in FactSage databases.

Similar to Figure 1, Figure 2 shows the fitted heat capacity and entropy functions of quaternary oxide $NaAlSi_3O_8$ and Figure 3 shows magnesium silicide (Mg_2Si). Our training set is the largest dataset to the best of our knowledge, and it encompasses a relatively large set of technologically important materials.

Entropy at 298 K

Entropy at 298 K is one of the twelve entropy models mentioned above. It is trained with 3190 compounds. Calculated mean absolute error (MAE) of a test set that has of 480 randomly selected compounds is 2.11 J/K per mole atom and root mean squared error (RMSE) is 3.09 J/K per mole atom. Uncertainty of experimentally determined S298 is around 1.1 J/K per mole atom according to the data provided by Kubaschewski [6].

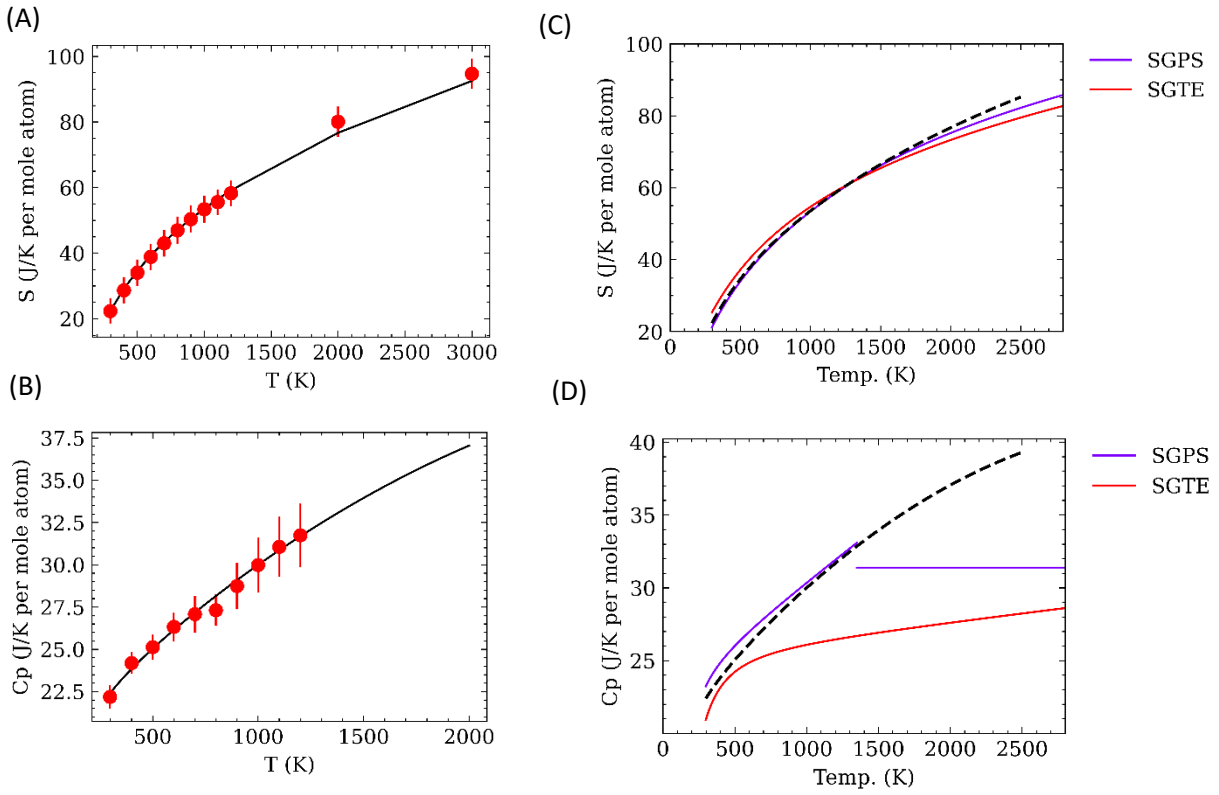

Figure 3. Heat capacity of Mg_2Si. (A): Estimated entropies. Red dots are the results of the GP regressions. Error bars are the uncertainties. Black line is the corresponding entropy curve of the fitted heat capacity function. (B): Similar to (A), red dots and error bars are the result of GP regressions. Heat capacity function is fitted on both entropy and heat capacity points and is the black curve. (C): Dashed line is the estimated entropy curve, i.e., same black curve in (A). Other functions are from the FactSage databases. (D): Estimated heat capacity and heat capacities in FactSage databases.

Formation Enthalpies and Corrections

Approximate DFT functionals result in systematic errors and a correction of DFT formation enthalpies is required. Even though Materials Project and OQMD are already applying such corrections, systematic errors are still observed. Thus, we apply additional corrections to the DFT values. Figure *1*Figure 4 shows a comparison between the Materials Project or OQMD calculated enthalpies and the corrected enthalpies of aiOQ and aiMP. The formation enthalpies for all compounds at room temperature are assumed to be the same as in the Materials Project or OQMD at 0 K if all constituting elements' ground states are the same at 0 K and 298 K and when the crystals do not contain functional groups.

If a constituting element's ground state changes between 0 K and 298 K or when the crystals contain functional groups, we follow a similar approach to Materials Project correction scheme [7]. But instead of a few hundred compounds, our training set contains 3547 compounds. In these cases, the formation enthalpies are corrected by element-specific corrections. We verify that 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. We train two different

models for Materials Project and OQMD datasets.

It should be noted that 10 kJ per mole atom is still significantly higher than the desired accuracies. Shifting the formation enthalpy of a compound by only 1 kcal per mole atom while keeping the liquid free energy curve the same might even lead a 500 K shift of congruent melting point. Thus, even though the thermodynamical data provided by the databases are reasonable, it is not expected to acquire accurate phase diagrams as is the case in other FactSage databases.

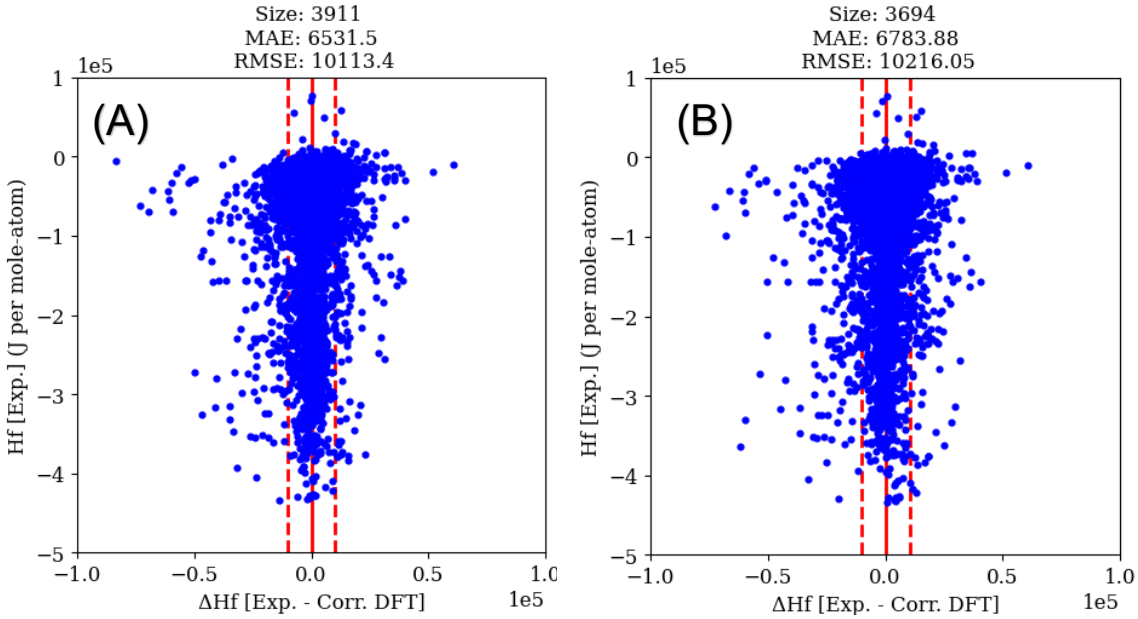

Figure 4. (A): Comparison between Materials Project calculated enthalpy of formation and the corrected enthalpy of formation used in aiMP. The solid red line represents the average error and the dashed red lines show the standard deviation. There exists a systemic error even though enthalpies provided by Materials Project already have a correction scheme that our correction scheme tries to fix. (B): Similar to (A), this shows the corrected enthalpies of formation for aiOQ vs the enthalpies of formation calculated with OQMD.

Density and Elastic properties

For all compounds the density is calculated from the relaxed unit cell volume given in Materials Project or OQMD. 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 are given in Materials Project.

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. Main reason for having such a set is the lack of data to benchmark and unfeasible training of ML models.

Additionally,

- Phases which have larger than 96 number of sites in their input cell are discarded.
- Phases which have amorphous tag are discarded.
- Phases with O and P from the Materials Project, which had been recalculated with R2SCAN functional, have been replaced with the previous GGA calculations due to worse agreement with experimental data for the compounds when using the R2SCAN fucntional
- Phases which have 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, first 50 with lowest formation enthalpy are included.
- If there are more than 7 elements exist in a compound, it is 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.

What is new?

Materials Project and OQMD databases are constantly being updated. Naturally, these changes are also applied to the aiMP/aiOQ v6.0. OQMD dataset is recently introduced.

Compared to the v5.0, v6.0 has a larger trainings set and more data from the Materials Project. The Materials Project has also introduced a new functional to their calculations: R2SCAN. The databases have been split into stable vs metastable databases to make calculations more efficient and easier.

Further Information

Please contact us via our ticketing system for customer support or via info@gtt-technologies.de if you need further information.

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.