
Further information on publisher’s website:
https://doi.org/10.1126/science.aad3000

Publisher’s copyright statement:
This is the author’s version of the work. It is posted here by permission of the AAAS for personal use, not for redistribution. The definitive version was published in Science on 25 Mar 2016: Vol. 351, Issue 6280, DOI: 10.1126/science.aad3000

Additional information:
Use policy

The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a link is made to the metadata record in DRO
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Please consult the full DRO policy for further details.
Reproducibility in density-functional theory calculations of solids


1Center for Molecular Modeling, Ghent University, Technologiepark 903, BE-9052 Zwijnaarde, Belgium,
2Peter Grünberg Institut and Institute for Advanced Simulation, Forschungszentrum Jülich and JARA, D-52425 Jülich, Germany,
3Department of Physics, Åbo Akademi, FI-20500 Turku, Finland,
4COMP/Department of Applied Physics, Aalto University School of Science, P.O. Box 11100, FI-00076 Aalto, Finland,
The widespread popularity of density-functional theory has given rise to a vast range of dedicated codes to predict molecular and crystalline properties. However, each code implements the formalism in a different way, raising questions on the reproducibility of such predictions. We report the results of a community-wide effort that compares 15 solid-state codes using 40 different potentials or basis set types, assessing the quality of the Perdew-Burke-
Ernzerhof equations of state for 71 elemental crystals. We conclude that predictions from recent codes and pseudopotentials agree very well, with pairwise differences comparable to those between different high-precision experiments. Older methods, on the other hand, show less precise agreement. Our benchmark provides a framework for users and developers to document the precision of new applications and methodological improvements.

Scientific results are expected to be reproducible. When the same study is repeated independently, it should reach the same conclusions. Nevertheless, some recent articles have shown that reproducibility is not self-evident. A widely resounding Science article (1) has dramatically demonstrated a lack of reproducibility of published psychology experiments. Although the more exact sciences are believed to perform better in this respect, concerns about reproducibility have emerged in these fields as well (2–4). The issue is of particular interest when computer programs are involved. Undocumented approximations or undetected bugs can lead to entirely wrong conclusions (5). In areas where academic codes compete with commercial software, unavailability of source code can hinder assessment of the relevance of the conclusions (6, 7).

Density-functional theory (DFT) calculations (8, 9) are a prominent example of a field that entirely depends on the development and appropriate use of complex software. With rigorous foundations in the quantum theory of matter, DFT describes the structure and properties of molecules and solids at the atomic scale. Over the years, many academic groups have developed implementations of DFT in computer codes, and several of these have been adopted by large user communities. Commercial alternatives are entering this area as well. At present, more than 15,000 papers are published each year that make use of DFT codes (10), with applications varying from metallurgy to drug design. Moreover, DFT calculations are nowadays used to build large databases (11, 12), or in multi-scale calculations where DFT codes are one part of
the tool chain (13, 14). It is no exaggeration to state that the precision of DFT codes underlies
the scientific credibility and reproducibility of a substantial fraction of the current natural and
engineering sciences, and thus reaches far beyond the traditional electronic-structure commu-
nity.

The main idea of density-functional theory is to solve the intractable many-particle Schrö-
dinger equation by replacing the complete electron wave function by the much simpler ground-
state electron density as fundamental variable. Although this reformulation is in principle exact,
it is not fully known how the interaction between individual electrons should be transformed.
As a result, the specific form of the unknown part of the interaction energy, the exchange-
correlation functional, has been the focus of many investigations, leading to a plethora of avail-
able functionals in both solid-state physics (15–19) and quantum chemistry (15, 20–23).

Once a particular exchange-correlation functional has been chosen, the mathematical problem
is completely specified as a set of so-called Kohn-Sham equations, whose solution yields
orbitals and energies from which the total electronic energy can be evaluated. A variety of such
numerical solution schemes have been implemented in different computer codes. Comparisons
of their performance are much less frequent or extensive than those of exchange-correlation
functionals, however (21, 24–29). One might reasonably expect that as they solve the same
equations they all produce similar answers for a given crystal structure, but a glance at the liter-
ature shows that this assumption is by no means always true. Fig. 1 demonstrates that even for
a well-studied material like silicon, predictions from different codes (the “precision”) vary by
the same order of magnitude as the deviation from the 0 K experimental value (26) (the “accu-
curacy”) (30). Because all codes depicted in Fig. 1 treat silicon at the same level of theory, using
the same exchange-correlation functional, they yield the same accuracy by definition. However,
the particular predictions vary from one code to another due to approximations unrelated to the
exchange-correlation functional. These approximations decrease the computational load, but
limit the precision.

What precision can we now achieve? Discussion of precision-related issues is uncommon in reports of solid-state DFT studies. The reproducibility of predictions is sometimes checked by cross-validation with other codes \((21, 24–28)\), but we are not aware of any systematic assessments of precision (also called “verification”) even though such studies would underpin confidence in practical DFT calculations. As a group of 69 code developers and expert users, we determine the error bar on energy-versus-volume \((E(V))\) predictions of elemental solids, running the same benchmark protocol with various DFT codes. Parameters of these equations of state (EOS), such as the lattice parameter or the bulk modulus, are commonly used for accuracy assessments \((15–19)\). By considering elemental solids, we establish a broad and comprehensive test for precision. Elemental solids exhibit a wide range of chemical environments, and constitute a reasonable first approximation to sample the huge compositional space of multicomponent systems. Our effort has resulted in 18 602 DFT calculations, which we aimed to execute with a rigorously determined precision. This exercise might seem simple, but each code tackles the Kohn-Sham equations and subsequent energy evaluation in its own way, requiring different approaches to deal with difficulties in different parts of the computational procedure.

**Kohn-Sham solution techniques**

The Kohn-Sham equations describe a many-electron system in terms of a density built from single-particle wave functions. By expressing these wave functions as a linear combination of predefined basis functions, the Kohn-Sham equations reduce to a matrix equation, which can in principle be solved exactly. Their solution should yield identical results irrespective of the form of the basis functions, provided the basis set is complete. However, achieving technical convergence of the complete Kohn-Sham problem is not feasible in practice. Consider silicon, whose electronic structure is schematically illustrated in Fig. 2. The Aufbau principle requires
first populating the lowest energy level, which for silicon is the 1s band. This is much lower in energy than the valence and conduction bands, and the localization of the orbitals close to the nuclei demands high spatial resolution. These core electrons do not contribute directly to chemical bonding, so they can be separated out and represented using a different basis, better suited to describe localized atomic-like states. Core orbitals may either be computed in an isolated atom environment and their effect on valence transferred unaltered to the crystal, or be relaxed self-consistently in the full crystal field. They can moreover be treated using a relativistic Hamiltonian, essential for core electrons in heavy atoms. Different relativistic schemes may lead to differences in the predicted $E(V)$ curves.

To stitch together a complete solution, the wave functions of the semi-core and valence electrons (2s 2p and 3s 3p, respectively, in the case of silicon) must be constructed to include the effect of orthogonality to the core electrons. This central problem can be solved in a number of different ways depending on the choice of numerical method. For methods based on plane-wave expansions or uniform real-space grids, the oscillatory behaviour cannot be accurately represented due to the limited spatial resolution. The need for unmanageably large basis sets can be mitigated by adding a carefully designed repulsive part to the Kohn-Sham potential, a so-called pseudopotential. This pseudopotential affects only a small region around the nuclei (grey zones in Fig. 2) and may conserve the core-region charge (norm-conserving pseudopotentials (31, 32)), giving rise to an analytically straightforward formalism, or break norm conservation by including a compensating augmentation charge (ultrasoft pseudopotentials (33)), allowing for smoother wavefunctions and hence smaller basis sets. Alternatively, the projector-augmented wave (PAW) approach defines an explicit transformation between the all-electron and pseudopotential wavefunctions using additional partial-wave basis functions (34, 35). This allows PAW codes to obtain good precision for small numbers of plane waves or large grid spacings, but choosing suitable partial-wave projectors is not trivial. We will refer to both pseu-
dopotential and PAW methods as pseudization approaches. In contrast, all-electron methods explicitly construct basis functions that are restricted to a specific energy range ((L)APW (36–39), LMTO (40)), or treat core and valence states on equal footing, e.g., by using numerical atomic-like orbitals (41, 42). Dealing with the full potential enables better precision, but inevitably increases the computation time. In these codes, the ambiguity in solving the Kohn-Sham problem shifts from the choice of the pseudization scheme to the choice of the basis functions. This choice lead to a variety of methods as well, which, despite solving the same Kohn-Sham equations, differ in many other details. Because each of these methods has its own fundamental advantages, it is highly desirable to achieve high precision for all of them.

**The $\Delta$ matrix**

The case study for silicon (Fig. 1) demonstrates that different approaches to the potential or basis functions may lead to noticeably varying predictions, even for straightforward properties like the lattice parameter. There is no absolute reference to compare these methods against; each approach has its own intricacies and approximations. To determine whether the same results can be obtained irrespective of the code or (pseudo)potential, we instead present a large-scale, pair-wise code comparison using the $\Delta$ gauge. This criterion was formulated by Lejaeghere et al. (26) to quantify differences between DFT-predicted $E(V)$ profiles in an unequivocal way. They proposed a benchmark set of 71 elemental crystals and defined for every element $i$ the quantity $\Delta_i$ as the root-mean-square difference between the equations of state of methods $a$ and $b$ over a $\pm 6\%$ interval around the equilibrium volume $V_{0,i}$. The calculated equations of state are lined up with respect to their minimum energy and compared in an interval symmetrical around
the average equilibrium volume (see Fig. 3):

\[
\Delta_i(a, b) = \sqrt{\frac{\int_{0.94V_0,i}^{1.06V_0,i} (E_{b,i}(V) - E_{a,i}(V))^2 \, dV}{0.12V_0,i}}
\]  

(1)

A comparison of \( \Delta_i \) values allows the expression of EOS differences as a single number, and a small \( \Delta_i \) automatically implies small deviations between equilibrium volumes, bulk moduli or any other EOS-derived observables as well. The overall difference \( \Delta \) between methods \( a \) and \( b \) is obtained by averaging \( \Delta_i \) over all 71 crystals in the benchmark set. Alternative definitions of \( \Delta \) have recently appeared as well \((27, 28)\), and essentially render the same information. In this work, we apply the original \( \Delta \) protocol to 40 DFT implementations of the PBE functional \((43)\). Appropriate numerical settings were determined for each method separately, ensuring converged results. In all calculations, valence and semicore electrons were treated on a scalar-relativistic level, as not all codes support spin-orbit coupling. This is not a limitation, as the aim is to compare codes to each other rather than to experiment. We do not elaborate much on speed and memory requirements, for which we refer to the documentation of the respective codes.

Fig. 4 presents an overview of the most important \( \Delta \) values categorized into all-electron, PAW, ultrasoft pseudopotential and norm-conserving pseudopotential methods. Approaches with a similar intrinsic precision are in this way clustered together. Both the full results and the most important numerical settings have been included in Tables S3–S42. A complete specification would have to include code defaults and hard-coded values, so a reasonable compromise was chosen. A full specification could be realized by recent endeavours towards full-output databases \((44, 45)\) or workflow scripting \((46, 47)\), but this is not yet available for several of the codes treated here. We have however tried to provide generation scripts for as many methods as possible \((48)\), and emphasize the need for such tools as an important future direction.
Comparing all-electron methods

Although the definition of $\Delta$ does not favour a particular reference, it is most instructive to first have a closer look at the $\Delta$ values with respect to all-electron methods (Fig. 4). They generally come at a computationally larger cost, but all-electron approaches to DFT are often considered to be a gold standard, as they implement the potential without pseudization. By comparing pseudopotential or PAW methods to all-electron codes, we therefore get an idea of the error bar on each pseudization scheme. The $\Delta$ values between different all-electron methods, on the other hand, reflect the remaining discrepancies, such as a different treatment of the scalar relativistic terms or small differences in numerical methods.

To gain some intuition into typical values of $\Delta$, we should first establish which values for $\Delta$ can be qualified as ‘small’, leading to results that can be considered equivalent. A first indication comes from converting differences between high-precision measurements of equation-of-state parameters into a $\Delta$ format. Comparing the high-quality experimental data of Holzapfel et al. for Cu, Ag and Au (49) to those of Kittel (50) and Knittle (51), for example, marks a small difference $\Delta_{exp}$ of 1.0 meV/atom. Since the average all-electron $\Delta$ for these materials is only 0.8 meV/atom, this implies that the precision of many DFT codes outperforms experimental precision. Secondly, we also break down the differences between codes in terms of commonly reported equation-of-state parameters. The 1.0 meV/atom maximum $\Delta$ between all-electron codes (Fig. 4, top) corresponds to an average volume deviation of 0.14 Å$^3$/atom (0.38 %) or a median deviation of 0.05 Å$^3$/atom (0.24 %) over the entire 71-element test set. For the bulk modulus the average deviation is 1.6 GPa (4.0 %) and the median deviation 0.8 GPa (1.6 %). Compared to the scatter on experimental values, amounting to up to 35 % for the bulk moduli of the rare-earth metals, for instance (52), these values are very small. The difference between equations of state obtained by independent all-electron codes is hence smaller than the spread.
between independent experimental equations of state. We conclude that, unless some elements deviate significantly from the overall trend, codes with a mutual $\Delta$ of 1 or even 2 meV/atom can be deemed to yield indistinguishable equations of state for all practical purposes.

The above-mentioned differences correspond to the best attainable precision for each all-electron code, using highly converged or ‘ultimate’ computational settings. It is important to realize, however, that particular choices of these settings may still slightly change the $\Delta$ values. Conversely, it is not always necessary to set such stringent requirements, as efficient codes are able to perform well with less-than-perfect settings. Nevertheless, the difference between default- and ‘ultimate’-precision equations of state may sometimes reach a few meV/atom (see Table S2). To eliminate the effect of numerical convergence altogether, we tested for the osmium crystal whether it was possible to obtain exactly the same result with different codes. Rather than aiming for the best representation of the ideal PBE results, as in the rest of this work, the goal was now to choose input settings as consistently as possible (using the same basis functions, grids and other parameters). Comparing four APW+lo calculations in this way yielded the results in Table 1. While numerical noise in various subroutines gives rise to fluctuations of only 0.02-0.04 meV/atom, the larger deviation of approximately 0.2 meV/atom in comparisons with respect to exciting can partly be attributed to a different scalar-relativistic treatment of the valence electrons in this code. Indeed, there is no single, universal method to account for the relativistic change of the electron mass in the kinetic energy. exciting uses the infinite-order regular approximation (IORA) (53), while the other three APW+lo codes use the Koelling-Harmon scheme (54). A third possibility is to use the zero-order regular approximation (ZORA), as was done in FHI-aims (42, 55) (see Tables S5–S7).
Comparing (pseudo)potential libraries

In comparison to all-electron codes, pseudization approaches are generally faster, as fewer states are considered and explicit construction and diagonalization of the Hamiltonian matrix is avoided. Among these, PAW and ultrasoft pseudopotentials require fewer basis functions than the norm-conserving variety, but advanced features such as linear response theory or hybrid functionals may sometimes not be available due to the increased complexity of the implementation. However, they all perform remarkably well in terms of precision when compared with all-electron results (see Fig. 4). For equations of state, the precision of current potentials is able to compete with that of all-electron methods, yielding $\Delta$ values of about 1 meV/atom, with a low approaching 0.3 meV/atom. This has not always been the case. As suggested by the example of silicon (Fig. 1), the available potentials have improved considerably over time. In Table 2 it can be seen that for several codes the $\Delta$ value is smaller for newer potential sets. Moreover, older potentials like the Troullier-Martins FHI98pp/ABINIT norm-conserving set, the Vdb2/DACapo ultrasoft set and the Vdb/CASTEP ultrasoft set all have a substantially larger $\Delta$ (Fig. 4). This evolution is evidence of internal quality control mechanisms used by developers of potentials in the past, as well as more recently, of additional efforts based on the $\Delta$ gauge (e.g., the JTH and SSSP potential libraries). The striking difference with the older potentials, even for the predefined structures in this relatively simple test set, provides a compelling argument to only use the most recent potential files of a given code.

In addition to a comparison with all-electron codes, it is also interesting to assess how different codes implement the same PAW or pseudopotential recipes. When both GPAW and ABINIT use the GPAW 0.9 PAW set, for example, they agree to within a $\Delta$ of 0.6 meV/atom. A similar correspondence is found for the Schlipf-Gygi 2015-01-24 ONCVPSP norm-conserving pseudopotentials (0.3 meV/atom between QUANTUM ESPRESSO and CASTEP), the GBRV 1.4
ultrasoft pseudopotentials (0.3 meV/atom between QUANTUM ESPRESSO and CASTEP) and the GBRV 1.2 set (0.7 meV/atom between PAW potentials in ABINIT and ultrasoft potentials in QUANTUM ESPRESSO). Here too, the small $\Delta$ values indicate a good agreement between codes. This agreement moreover encompasses varying degrees of numerical convergence, differences in the numerical implementation of the particular potentials and computational differences beyond the pseudization scheme, where the latter are expected to be of the same order of magnitude or smaller than the differences between all-electron codes (1 meV/atom at most).

**Conclusions and outlook**

Solid-state DFT codes have evolved tremendously. The change from small and personal codes to widespread general-purpose packages has pushed developers to aim for the best possible precision. Whereas DFT-PBE literature on the lattice parameter of silicon displayed a spread of 0.05 Å in the past, the most recent versions of the implementations discussed here agree on this value by 0.01 Å (see Fig. 1 and Tables S3–S42). By comparing codes on a more detailed level using the $\Delta$ gauge, we have indeed found most recent methods to yield nearly indistinguishable equations of state, with the remaining error bar comparable to that between different high-precision experiments. This underpins the validity of recent DFT EOS results and ensures that correctly converged calculations yield reliable predictions. The message moreover impacts all of the multi-disciplinary set of fields that build upon DFT results, ranging from physical to biological sciences.

In spite of the absence of one absolute reference code, we were able to improve and demonstrate the reproducibility of DFT results by means of a pair-wise comparison of a wide range of codes and methods. Any new methodology development can now verify whether it can reach the same precision, and new DFT applications can show to have used a method and/or potentials that were screened in that way. The data generated in the framework of this paper serve as
a crucial enabler for such a reproducibility-driven paradigm shift, and future updates of available $\Delta$ values will be presented at http://molmod.ugent.be/deltacodesdft. The reproducibility of reported results also provides a sound basis for further improvement of the accuracy of DFT, i.e. the investigation of new DFT functionals, or for the development of new computational approaches. This work might therefore speed up methodological advances in solid-state DFT substantially.

There is scope for future work to check the reproducibility of different codes even further. This might consider larger benchmark sets (describing different atomic environments per element), other functionals, an exhaustive comparison of different relativistic treatments, and a more detailed account of computational differences (using databases or scripts, for example). The precision of band gaps, magnetic anisotropies and other non-EOS properties would also be of interest. However, the current investigation of equation-of-state parameters provides the most important pass/fail test to the quality of different implementations of Kohn-Sham theory. A method that is not able to reach an acceptable precision with respect to the equations of state of the elemental crystals, will likely not fulfill even more stringent demands.

References and Notes


30. The studies of accuracy and precision for DFT calculations are often referred to as validation and verification (V&V), respectively.


82. http://pwpaw.wfu.edu/.


**Acknowledgments**

The present research benefited from financial support from the Research Board of Ghent University (BOF), the Fond de la Recherche Scientifique de Belgique (FRS-FNRS) through PDR Grants T.0238.13 - AIXPHO and T.1031.14 - HiT4FiT, the Communauté française de Belgique through the BATTAB project (ARC 14/19-057), NSF Grant DMR-14-08838, the Swedish Research Council (VR), the Knut and Allic Wallenberg (KAW) foundation, the Fund for Scientific Research–Flanders (FWO) (project number G0E0116N) and the US Department of Energy through grant DOE-BES DE-SC0008938.

The work by N.A.W.H. was supported by NSF grant no. DMR-1105485. J.A.F.-L. acknowledges financial support from the EU’s 7th Framework Marie-Curie Scholarship Program within the “ExMaMa” Project (Project No. 329386). I.D.M., O.E., O.G., D.I., Y.O.K., I.L.M.L., and L.N. acknowledge support from eSSENCE. T.B. was supported by the Academy of Finland.
through Grant No. 263416 and the COMP Centre of Excellence. C.D., A.G., and S.L. acknowledge support from the DFG and the Einstein Foundation, Berlin. A.D.C., S.d.G. and E.K. acknowledge support from the Italian MIUR through PRIN 2010-11 (Registration number 20105ZZTSE_005). P.J.H., D.B.J. and M.I.J.P. are grateful for financial support by EPSRC under the UKCP Grant EP/K013564/1. C.J.P. and J.R.Y. acknowledge support from the Collaborative Computational Project for NMR Crystallography under EPSRC grant EP/J010510/1. W.P. acknowledges funding by FWO. D.J. is grateful for financial support by EPSRC under grant EP/J017639/1. S. Saha acknowledges the support from the Swiss National Science Foundation (SNF). G.-M.R. is thankful for personal financial support from FRS-FNRS. The work by I.E.C. and N.M. was supported by the NCCR MARVEL of the Swiss National Science Foundation. G.K. and P.B. acknowledge support by the Austrian Science Fund (FWF), Project SFB-F41 (ViCoM). S.C. acknowledges financial support from OCAS NV by an OCAS-endowed chair at Ghent University.

Computational resources were provided by: for the Ghent University contributors, the Stevin Supercomputer Infrastructure at Ghent University, funded by Ghent University, the Hercules Foundation, and the Flemish Government (EWI Department); for the Université catholique de Louvain contributors, the Tier-1 supercomputer of the Fédération Wallonie-Bruxelles, infrastructure funded by the Walloon Region under the grant agreement no. 1117545; the CISM/UCL supercomputing facilities and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie-Bruxelles (CÉCI) funded by the FRS-FNRS under convention 2.5020.11; for the CASTEP calculations, the STFC Scientific Computing Department’s SCARF cluster; for the Basel University and EPFL contributors, the Swiss National Supercomputing Center (CSCS) in Lugano; for the RSPt contributors, Finland’s IT Centre for Science (CSC).

K.L. and F.T. would also like to thank Chandler Becker for instructive discussions on the comparison of atomic-scale simulations.

Commercial software is identified to specify procedures. Such identification does not imply recommendation by the National Institute of Standards and Technology.

ASE scripts (46) for several of the codes are available online (48). All data are listed in Tables S3–S42.

Supplementary Materials

Materials and Methods

Fig. S1

Tables S1 to S42

References (48, 66–115)
Fig. 1. Historical evolution of the predicted equilibrium lattice parameter for silicon. All data points represent calculations within the DFT Perdew-Burke-Ernzerhof (PBE) framework. Values from literature \((15, 16, 18, 56–65)\) are compared both to the predictions of different codes within this paper (data points from 2016 and inset; older methods or calculations with lower numerical settings have been depicted by empty symbols) and to the experimental value extrapolated to 0 K and corrected for zero-point effects (red line) \((26)\). The concepts of precision and accuracy (see text) are illustrated graphically.

Fig. 2. Electronic states in solid silicon. The valence states are delocalized over the solid (green line), as the wave functions overlap from one atom to the next. The lowest-energy 1s state (red) is at an energy two orders of magnitude lower than the valence states, and is strongly localized near the nucleus with no overlap between the atoms. The grey regions around the atoms indicate approximately where the wavefunction, density and potential are smoothed in pseudized methods.

Fig. 3. Graphical representation of the \(\Delta\) gauge. The black line depicts the quadratic energy difference between two equations of state, and \(\Delta_i\) corresponds to the root-means-square average. This is demonstrated by the shaded area, which is equally large above and below the \(\Delta_i^2\) line.

Fig. 4. \(\Delta\)-values between the most important DFT methods considered (in meV/atom). Comparison of all-electron (AE), PAW, ultrasoft (USPP) and norm-conserving pseudopotential codes (NCPP) to all-electron results, listed in alphabetical order per category. The tags stand for code, code/specification (AE) or potential set/code (PAW/USPP/NCPP), and are specified in full in Tables S3–S42. The colour code ranges from green over yellow to red (small to large \(\Delta\) values). The mixed potential set SSSP was added to the ultrasoft category, in agreement with its prevalent potential type. Both the code settings and the DFT-predicted equation-of-state parameters behind these numbers have been included in Tables S3–S42, and
a full $\Delta$ matrix for all methods mentioned in this article is available in Fig. S1.

**Table 1. Agreement between osmium crystal predictions at nearly identical settings. (Top)**

$$\Delta_i$$ values for the osmium crystal (in meV/atom) when four APW+lo calculations tried to mimic the same settings as well as possible. These settings are therefore different from the ones used for Fig. 4 and reported in Tables S3–S4, S8 and S15. **(Bottom)** The corresponding equilibrium volumes $V_0$, bulk moduli $B_0$ and bulk modulus derivatives $B_1$.

**Table 2. Precision evolution of PAW and pseudopotential sets over time.** The $\Delta$-values are expressed as an average over the all-electron methods (in meV/atom) and are listed chronologically per code. Both the corresponding code settings and the DFT-predicted equation-of-state parameters have been listed in Tables S17, S19–S26, S30–S31 and S33. The most recent potentials are the ones used to generate the data shown in Fig. 4.