Journal of Materials Chemistry C

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: J. Mater. Chem. C, 2025, 13, 18197

Accelerating the discovery of high-performance nonlinear optical materials using active learning and high-throughput screening†

Victor Trinquet, *[a] Matthew L. Evans [ab] and Gian-Marco Rignanese *[acd]

Received 28th March 2025,
Accepted 17th July 2025

DOI: 10.1039/d5tc01335f

rsc.li/materials-c

Due to their abundant use in all-solid-state lasers, nonlinear optical (NLO) crystals are needed for many applications across diverse fields such as medicine and communication. However, because of conflicting requirements, the design of suitable inorganic crystals with strong second-harmonic generation (SHG) has proven to be challenging for both experimentalists and computational scientists. In this work, we leverage a data-driven approach to accelerate the search for high-performance NLO materials. We construct an extensive pool of candidates using databases within the OPTIMADE federation and employ an active learning strategy to gather optimal data while iteratively improving a machine learning model. The result is a publicly accessible dataset of ~2200 computed SHG tensors using density-functional perturbation theory. We further assess the performance of machine learning models on SHG prediction and introduce a multi-fidelity correction-learning scheme to refine data accuracy. This study represents a significant step towards data-driven materials discovery in the NLO field and demonstrates how new materials can be screened in an automated fashion.

1 Introduction

Thanks to their frequency-conversion properties, nonlinear optical (NLO) materials play a significant role in modern optoelectronics.[1] Their ability to produce coherent light by up- or down-converting incident electromagnetic waves has found applications in a variety of fields, from laser technologies and optical communication to biomedical imaging and quantum information processing.[2–5] As is often the case with functional materials, a good NLO compound needs to meet several requirements such that it turns out to be a multi-objective optimisation. This ends up limiting the number of efficient materials, especially in the deep ultraviolet (DUV), the mid-, and the far-infrared (IR) ranges.[6,7] It is thus of interest to accelerate the discovery of novel NLO materials, for both academic and industrial purposes.

As things stand, experimental studies lack the speed and cost-efficiency to freely consider the whole compositional and structural space. For this reason, computational methods are increasingly being used to navigate the almost endless possibilities.[8] In practice, the search for NLO materials is translated into a search for appropriate compounds displaying strong second-harmonic generation (SHG), which enables a doubling of the incident frequency. Using density-functional theory (DFT), the SHG tensor can be calculated and investigated with respect to the chemistry and structure of a given compound. Many studies have thus focused on the efficient design of novel NLO crystals.[9–11] Another approach relies on high-throughput screenings of existing databases to identify promising materials.[12–15] The latter can then be used to suggest other candidates and to investigate unexplored families of compounds. However, large open-access databases do not readily provide the SHG tensors.[16,17] Other basic properties are usually used to restrict the DFT computations of SHG tensors to stable non-centrosymmetric (NCS) crystals with an electronic band gap in the range of interest. Although this procedure has led to the emergence of a few datasets with SHG information, the domain is definitely lacking significant NLO datasets that could be used for efficient screening or materials informatics.[12,18,19] To address this issue, Xie *et al.*[20] computed 1500 SHG tensors of stable NCS semiconductors from the Materials Project (MP)[16] in 2023. Combined with 900 materials generated *via* an evolutionary algorithm, this dataset is a first step towards big data in the NLO field. In 2024, Wang *et al.*[14] also performed a screening of the MP involving the computation of ~2400 SHG tensors.

[a] UCLouvain, Institut de la Matière Condensée et des Nanosciences (IMCN), Chemin des Étoiles 8, Louvain-la-Neuve 1348, Belgium. E-mail: victor.trinquet@uclouvain.be, gian-marco.rignanese@uclouvain.be
[b] Matgenix SRL, 185 Rue Armand Bury, 6534 Gozée, Belgium
[c] WEL Research Institute, Avenue Pasteur 6, 1300 Wavre, Belgium
[d] School of Materials Science and Engineering, Northwestern Polytechnical University, No. 127 Youyi West Road, Xi'an 710072, Shaanxi, China

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d5tc01335f

Recent years have seen a significant increase in the amount of available data related to materials properties. Existing experimental and computational databases are continuously growing while new actors and initiatives appear.[16,17,21–23] This trend has been accelerating with the emergence of data-driven and machine learning approaches that are able to generate hypothetical compounds, many of which are predicted to be stable, to some definition.[24–26] Although this growth in data presents new opportunities for materials discovery, it also provides new challenges that require rational screening methods to help efficiently allocate experimental resources within this growing design space. This is where data standardisation and federation can play an important role. The OPTIMADE consortium[27,28] consists of several leading crystal structure database providers and datasets that have agreed upon a common data format and query language, enabling seamless access to over 60 million structures across 30 decentralised databases. Several of these databases are targeted towards assessing materials stability, typically using DFT, providing a fruitful and growing pool for screening compounds with potentially exemplary properties in order to prioritise costly synthesis attempts.

In this work, we aim at propelling the NLO field into the era of big(ger) data while addressing the above challenge when generating and navigating the candidates design space. The end goal is the discovery of NLO bulk inorganic crystals with strong second-harmonic conversion. In practice, the search is translated into a multi-objective optimisation involving conflicting physical quantities, namely the SHG tensor and the band gap. For a given strength of SHG, maximising the band gap ensures a broad transparency window while promoting higher laser damage thresholds, an important practical consideration. By leveraging the common application programming interface (API) designed by the OPTIMADE federation, we easily build a large pool of candidate materials that will continue to grow as more structures and databases come online. This design space is then searched for good NLO materials using a cheap machine learning (ML) model trained on an existing dataset of SHG tensors. Since DFT computations of SHG tensors are resource intensive and the size of the initial dataset is limited, we adopt an active learning (AL) procedure for the ''training–predicting–selecting–computing'' steps in the data acquisition process. This allows us to efficiently target promising materials in this large search pool, whether they are interesting for the combined SHG-band gap optimisation, or for improving the accuracy of the machine learning model.

This paper first describes the computational workflow for computing the static SHG tensors, the details of the active learning procedure and the candidate pool generation. The end result is a dataset of $\sim$2200 static SHG tensors, which is made publicly available on the Materials Cloud Archive, itself accessible *via* an OPTIMADE API.[29,30] Thanks to this new dataset, we explore the performance of various ML algorithms on the present SHG task and we investigate a multi-fidelity correction-learning scheme to alleviate the inherent limitation of our data. Finally, we list the most promising materials uncovered in our dataset and look onward to the continued screening of large databases of hypothetical materials.

2 Methods

2.1 First-principles calculations

The quantity of interest in this work is the third-rank tensor responsible for second-harmonic generation.[1,31,32] This nonlinear optical phenomenon naturally appears in the framework of perturbation theory when the macroscopic polarisation, **P**, is expressed as a power series of the incident electric field, **E**, such that

$$P_i = \varepsilon_0 \sum_j \chi^{(1)}_{ij} E_j + \varepsilon_0 \sum_{jk} \chi^{(2)}_{ijk} E_j E_k + \text{higher order terms}, \quad (1)$$

with ε_0, the vacuum permittivity, and $\chi^{(1)}$, the linear susceptibility. The nonlinear susceptibility, $\chi^{(2)}$, is responsible for SHG in the case of two incident fields at the same frequency. By convention, this tensor is halved and is commonly referred to as the SHG tensor, **d**. By symmetry, the Voigt form can be adopted, thereby reducing it to a 3×6 second-rank tensor. It is important to note that only NCS compounds can display non-zero components of the SHG tensor. To facilitate visualisation and comparison across different materials, an effective scalar coefficient, d_{KP}, can be derived following the Kurtz–Perry (KP) powder method.[33]

In the present work, the open-source first-principles software ABINIT is used to compute the static limit of the SHG tensor in the framework of density-functional perturbation theory (DFPT).[34–38] The exchange–correlation energy is modelled using the local-density approximation (LDA) by using the optimised norm-conserving pseudopotentials from the PseudoDojo (scalar relativistic v0.4.1), which also provides cutoff values (''standard'' accuracy with hint ''normal'').[39,40] From the latter, the plane-wave energy cutoff is set based on the hardest element for each compound. The total energy is converged within 1×10^{-22} Ha during the self-consistent field cycles while the convergence tolerance on the wavefunction and the potential residual are respectively set to 1×10^{-22} Ha^2 and 1×10^{-22} Ha in the response function calculations. The Brillouin zone is sampled with a density of 3000 points per reciprocal atom (kppa) as it constitutes a reasonable balance between computational convergence and efficiency (see Section A1, ESI†). This sampling respects the symmetry of the system.

These high-throughput calculations are performed using the ShgFlowMaker class implemented in the atomate2 Python package[41] as jobflow workflows.[42] Since it defaults to the aforementioned *k*-point grid, only the type of pseudopotentials must be explicitly set to reproduce our results. This workflow is similar to the one presented in Trinquet *et al.*[43] apart from the pseudopotentials version and a revised algorithm to generate the *k*-points. Combined with the FireWorks workflow manager and the MongoDB database engine, this tool handles calculation submission and retrieval of the results.[44] Sometimes, the SHG tensor requires a rotation in order for its components to

match the conventional form set by the IEEE.[45,46] Both raw and post-processed tensors are made available.

The materials exhibiting a good balance between the KP coefficient and the band gap are selected for additional calculations to further refine their SHG tensors. Higher accuracy can indeed be achieved by including a rigid shift of the conduction bands. Up to this point, the band gaps were directly taken from the source databases at the Perdew–Burke–Ernzerhof generalised-gradient approximation (GGA-PBE) level.[47,48] In order to obtain the values of the scissor shifts, the band gaps of the crystals of interest are computed at two different levels. First, the LDA band gaps are computed thanks to the ABINIT BandStructureMaker class of atomate2 using the same set of pseudopotentials as the SHG calculations. The number of divisions to sample the smallest segment of the high-symmetry path is set to 15. The band gap is then taken as the lowest gap value across both the self-consistent field (SCF) and the non-SCF calculations. Second, the higher accuracy band gap is computed using the Heyd–Scuseria–Ernzerhof (HSE) hybrid functional[49,50] as implemented in VASP with projector augmented wave pseudopotentials[51–53] (PBE_64). A first SCF step is performed with PBE beforehand to facilitate the convergence. This process is automated by linking the HSEBSMaker class to the VASP StaticMaker class of atomate2. The electronic self-consistent loops are considered converged when a difference in energy lower than 1×10^{-6} eV is reached. The self-interaction energy is corrected using element-specific Hubbard U values[54] recommended by the MP.[55] In both the initial PBE and the LDA SCF steps, the Brillouin zone is sampled using a uniform grid with a density of 1500 points per reciprocal atom. The difference between the HSE and LDA gaps provides a scissor shift to refine the SHG tensors. These band gap corrections can be given to the ShgFlowMaker class to correct the DFPT computation.

It was decided not to perform any structural optimisation of the crystals, since the source databases already performed such relaxations at the GGA-PBE level using compatible settings. Despite this choice, and despite the adoption of the LDA, the most basic representation of the exchange–correlation energy in DFT, this approach to computing SHG coefficients has previously shown good results in ranking the SHG strength of materials with respect to their experimental values.[43] Moreover, the perturbative DFPT method used here incorporates local-field effects, which tend to slightly diminish the SHG strength as shown by time-dependent density-functional theory calculations;[56,57] these effects are excluded in typical sum-over-states calculations in the independent-particle approximation. Good experimental agreement has been shown when excitonic contributions can be included,[56,57] however, their inclusion is currently intractable for high-throughput studies.

2.2 Active learning

Similarly to our previous work, an active learning loop is adopted to optimally guide the acquisition of new data.[58] In practice, cheap machine learning predictions of the KP coefficient are used to select materials whose SHG tensor will be computed using the more expensive DFPT method, thus extending the available SHG dataset for training. While the end-goal is the discovery of new materials boasting high SHG coefficients for a given band gap, it is still of interest to spend computing resources on suboptimal compounds, provided that their addition in the training set significantly improves the performance of the surrogate model. In this work, the predictions and their corresponding uncertainties come from the average and standard deviation of predictions from a MODNet ensemble, *i.e.*, an ensemble of neural-networks[59–61] (see Section A2, ESI†). Since the methodology and choice of ML model are similar to Trinquet *et al.*,[58] only the differences are described hereafter. Fig. 1 illustrates the global methodology. In contrast to the case of the refractive index, we are not aware of any effective quantity whose maximisation could replace the optimisation of the (E_g, d_{KP}) Pareto front; instead, here we sample explicitly from the Pareto front of our candidates.

At each iteration, a MODNet ensemble is trained on T_i, the training set at the ith iteration of the AL process. This ML model yields a prediction of the KP coefficient, $p_i(d_{KP}|x) \sim \mathcal{N}(\mu_{d_{KP},i}, \sigma_{d_{KP},i})$ for a material x with mean ensemble model prediction $\mu_{d_{KP},i}(x)$ and uncertainty $\sigma_{d_{KP},i}(x)$. This allows us to define an upper bound for the target for each material as follows:

$$d_{U,i}(x) = \mu_{d_{KP},i}(x) + \lambda \cdot \sigma_{d_{KP},i}(x), \tag{2}$$

where the balance between exploration and exploitation is determined by the dimensionless parameter, λ. In order to diversify the selected compounds, the following regimes can be adopted:

$0 \rightarrow$ highest mean (uncertainty-agnostic exploitation)

$-1 \rightarrow$ highest mean with lowest uncertainty

$1 \rightarrow$ highest mean with highest uncertainty

$\lambda_{cal.} \rightarrow$ highest mean with high calibrated uncertainty

$\infty \rightarrow$ highest uncertainty (exploration)

The calibration factor, $\lambda_{cal.}$, is obtained by minimising the miscalibration area on a hold-out set and is then averaged over a 5-fold splits. It was found to lie consistently between 1.2 and 1.5 across all AL cycles.

The compounds are selected based on the following acquisition function:

$$\alpha_i(x) = \begin{cases} 1 & \text{if } x \in \mathscr{F}_{U,i} \\ 0 & \text{else} \end{cases} \tag{3}$$

where $\mathscr{F}_{U,i}$ is the Pareto front of the (E_g, $d_{U,i}$) distribution built from the entire candidate pool of materials, $\mathscr{P}$. This front is determined purely geometrically working from high to low band gap, after removing candidates with more than 50 atoms in the primitive unit cell. Since $d_{U,i}$ can be defined according to several regimes, the Pareto front for each λ regime, $\mathscr{F}_{U,i}(\lambda)$, is

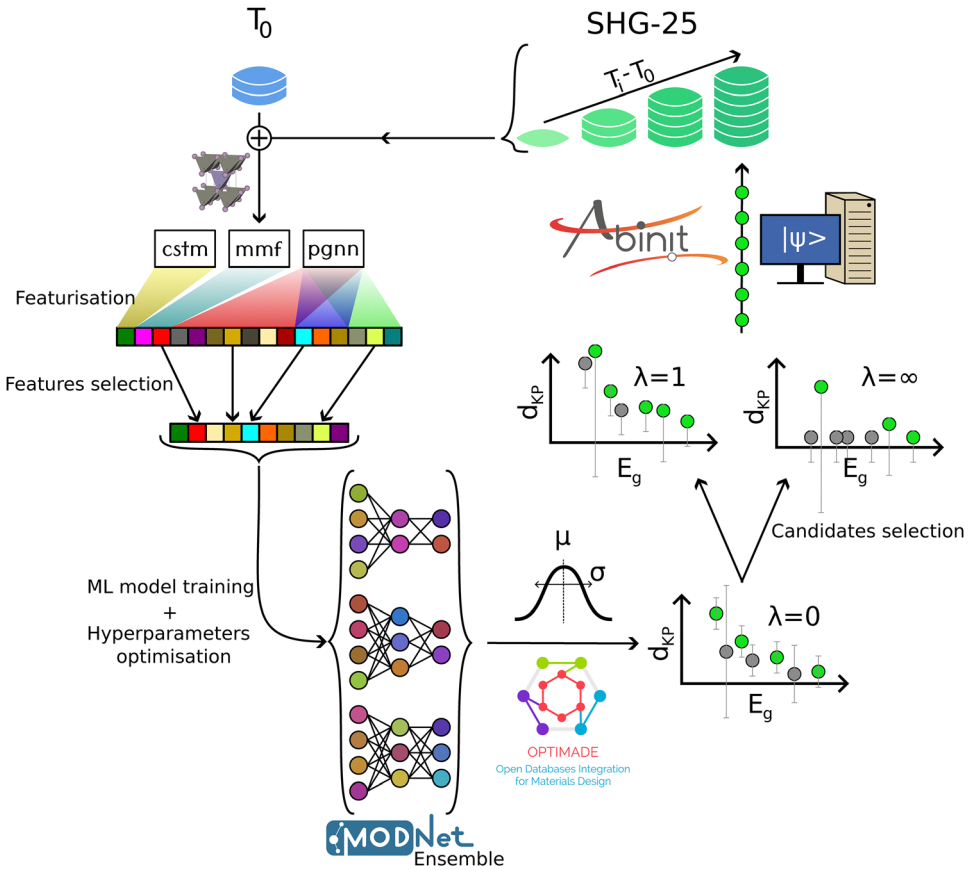

Fig. 1 Schematic of the active learning process. The initial training dataset, T_0, is featurised into physical and chemical descriptors using the matminer (mmf) and pGNN Python packages. The band gap of the source databases and the predicted refractive index with its uncertainty are also considered as additional custom features (cstm). The most relevant features are then ranked and, starting from the top of this ranking, the number of selected features is optimised during the hyperparameters optimisation. Once the MODNet ensemble has been trained, it predicts the KP coefficient along with a given uncertainty for the whole pool of candidate materials that was queried with the OPTIMADE API. From these predictions and uncertainties, the different exploitation and exploration regimes of the selection algorithm determine the Pareto front in the (E_g, $\mu_{d_{KP}} + \lambda\sigma_{d_{KP}}$) space (in green). The static SHG tensors of those promising entries are then computed within DFPT and added to the new (and initially empty) SHG-25 dataset. The latter is then combined with T_0 and a new AL cycle can begin.

found and they are all merged to form the selected subset, $\mathscr{F}_{\forall,i}$. If the latter is not large enough, it is removed from the distribution and extended by the front of this new distribution. This acquisition function effectively classifies $\mathscr{P}$ at each AL cycle. This combination of exploration and exploitation along with the selection balancing the band gap and the KP coefficient ensures that the data acquisition remains as free as possible from any unwanted bias introduced by the inaccuracy of the ML model. On the contrary, the latter should be affected positively by the new data. After running the DFPT calculations on $\mathscr{F}_{\forall,i}$, a compound is flagged as an outlier if its d_{KP} is greater than 170 pm V^{-1} or if its static refractive index is greater than 20. The cleaned DFPT results (without outliers) are then added to the training set for the next iteration:

$$T_{i+1} = T_i \cup \mathscr{F}_{\mathrm{DFPT},i} \quad \text{with} \quad \mathscr{F}_{\mathrm{DFPT},i} \in \mathscr{F}_{\forall,i}. \qquad (4)$$

A new MODNet model is then trained from scratch on T_{i+1}, thus initiating a new iteration. This process can be stopped based on arbitrary criteria involving the model accuracy, the size of the training set, the coverage of the materials, their performance, or the available computing resources.

2.3 Training and candidate data

The dataset from Trinquet *et al.*[43] serves as the initial training set, T_0. It comprises 579 SHG tensors of inorganic semiconductors computed with ABINIT using the DFPT procedure outlined in Section 2.1. It should be noted that these calculations used an older set of pseudopotentials than presented here.

The MODNet model, feature selection algorithm, hyperparameters optimisation and training procedure follow those described in Trinquet *et al.*[58] (see Section A2, ESI†). A first set of ~200 physical and chemical descriptors was generated using the matminer Python package *via* the Matminer2024FastFeaturizer preset implemented in MODNet v0.4.3.[62] A second set of ~1000 features, referred hereafter as pGNN, is derived using the rogeriog/pGNN Python package[63] and appended to the first ~200 features (see Section A2, ESI†). Moreover, the final MODNet model of Trinquet *et al.*[58] predicts refractive indices and their uncertainties, which are appended to the set of

features, along with the band gap found in the source databases (computed with PBE). These additional descriptors were considered due to the known relationship they share with the SHG strength.[64] It is important to note that the actual set of features used in the AL loop was not fixed from the start as these iterations were refined over several months. Additional descriptors were tested during this process and added to the feature set, if they were deemed useful, as illustrated in Fig. A4 (ESI†). Since the resulting dataset of Trinquet *et al.*[58] was built to target the refractive index – band gap Pareto front, the SHG coefficient of each of its constituent NCS materials was computed, independently of the AL selection scheme.

Two source databases are considered to form the initial search pool, $\mathscr{P}$. The first one is the Materials Project (MP) with ∼160k materials (v2023.11.1),[16] resulting from DFT relaxations of primarily experimentally determined crystal structures from the inorganic crystal structure database (ICSD).[22] Using a combination of the MP API and their corresponding OPTIMADE API, the MP was filtered for NCS inorganic crystal structures possessing a PBE-computed band gap greater than 0.05 eV and a distance from the MP convex hull (by the latest mixed GGA+*U*/mGGA workflow) lower than 50 meV per atom. The resulting set of compounds is further reduced by excluding any lanthanide- or actinide-containing compounds, effectively reducing the MP to a subset of ∼13.5k relevant crystal structures relaxed with GGA-PBE. The second database is Alexandria[17,65] with its ∼4.5 M PBE-relaxed structures (v2023.12.29). Thanks to the OPTIMADE API, this vast number of entries is filtered for the same criteria as the MP. This query added ∼30.6k relevant structures to the pool of trial materials available for the AL process.

Duplicates across these two databases were removed by combining entries that share the same composition and space group, in which case the MP entry was preferred. The final candidate pool, $\mathscr{P}$, spans ∼33.5k NCS stable semiconductors.

It should be noted at this stage that both MP and Alexandria now contain additional entries matching our criteria which were not present at the initiation of our AL procedure. Additionally, new databases have been made available through OPTIMADE, such as the GNoME dataset,[24] which contains several hundred thousand hypothetically stable compounds. This study could thus be viewed as an intermediate step of a broader screening which will continue as new hypothetical compounds are suggested, and can act to prioritise experimental resources towards verifying the computed structures.

2.4 Benchmarking ML models for SHG

The second part of this work investigates the performance of various ML models on the prediction of the computed d_{KP} coefficient. Both T_0 and the newly acquired SHG tensors are considered. The dataset is cleaned by removing any outliers or duplicates found by the default StructureMatcher of the pymatgen Python package.[66] After removing the materials that fall abnormally far away from the data distribution (indicating, *e.g.*, a convergence issue), ∼2600 instances remain.

The different models were benchmarked on a holdout set of 250 compounds sampled such that the distribution of the target d_{KP} values matched that of the full dataset. Three other holdout sets were benchmarked (one different size and two random sets), with additional results reported in the complementary GitHub repository. In the presentation of these benchmarks, we will focus on the former dataset, as shown in Fig. A2 (ESI†), as we believe it to be the most representative. Given the large range of target values and the clear bias of the dataset towards low values, this procedure allows for a more robust comparison than a single test set while being less computationally intensive than full cross-validation.

When needed for hyperparameter optimisation, a validation set was sampled from the training set using the same algorithm that was used to generate the test set. The resulting set of hyperparameters was then adopted for training the model on the whole training set before assessing it on the holdout sets. For descriptor-based models, both the Matminer2024FastFeaturizer preset and the pGNN features[63] were considered. Three sets of features are derived: mmf with only the former, pgnn with only the latter, and mmf_pgnn merging both of them.

Several classes of ML models were investigated, from simple feed-forward neural networks like MODNet,[59,60] to tree-based methods (extra trees and LGBM),[67–69] graph neural networks (co(N)GN,[70] MEGNet,[71] TensorNet[72] for scalar predictions and Matten[73] for full tensor predictions) as well as several commercial (GPT-4o, Claude Sonnet 3.7) and open (DARWIN 1.5[74]) large language models (LLMs). A description of each model and any specifics of the training procedure or hyperparameter optimisation for each model are provided in Section A2 (ESI†).

Model performance was assessed using standard metrics: MAE, RMSE, R^2, and most relevant for screening studies, Spearman's rank correlation coefficient. In addition to these simple metrics, enrichment factors and discovery curves were computed for each model and holdout set. An enrichment factor (EF) defined at a given percentage, say EF (10%), corresponds to the reduction in the number of oracle evaluations (in this case DFT calculations) required to find the top 10% of materials. For example, for a set of 100 candidate materials, if following the model's predicted ranking would allow the top 10% to be found after 20 evaluations, the EF (10%) would be 5, out of at theoretical maximum of 10, or to compare across different thresholds, this can be normalised to 0.5. This metric is particularly important given the skewed nature of our dataset; a model could achieve reasonable performance in the low-SHG regime without being an effective discriminator of exemplary materials and *vice versa*. Discovery curves provide a generalisation of the enrichment factor, by spanning the entire range of percentiles; they are conceptually similar to receiver-operating characteristic (ROC) curves, extended to a global ranking rather than binary classification at different probability thresholds.

2.5 Multi-fidelity correction-learning

Few computational SHG datasets exist in the literature, and even fewer are publicly available.[12,18–20] However, a common

trait that most of them share is the adoption of scissor correction, which is a rigid shift of the conduction bands to match high accuracy band gaps obtained with hybrid functionals. By artificially opening the band gap, one can alleviate the overestimation of the SHG components caused by the usual underestimation of the band gap by local and semi-local functionals of the exchange–correlation energy. Contrary to those datasets, our SHG DFPT calculations in the AL scheme do not include scissor shifts. One could argue that this choice limits the impact of our new dataset, which would be true if the relative ranking of the materials were very different when considering a scissor shift, as any high-throughput screening involving our data would then be meaningless.

To address this concern, we correct a subset of the compounds of the final AL dataset (including the initial training set) and show that this is not the case, *i.e.*, that the uncorrected SHG coefficients are sufficient for screening. Since the computational cost of the more accurate HSE band gaps is non-negligible, it is sensible that only the optimal materials in the (E_g, d_{KP}) space benefit from this correction.

The compounds to be scissor-corrected were selected by recursively determining the (E_g, d_{KP}) Pareto front and choosing the constituent compounds without replacement. In accordance with the available computational resources, this process was repeated until $\sim$1000 compounds were obtained. As described in Section 2.1, both their ABINIT LDA and their VASP HSE gaps are computed to derive scissor shifts, which are used in subsequent DFPT SHG computations. Compounds with an HSE gap lower than 1 eV are discarded. In addition, the selected entries from T_0 are also computed at the LDA level to ensure that their KP coefficient is consistent with the ones of SHG-25.

Since SHG tensors at both the LDA and the "HSE" level are now available, machine learning algorithms can be used to leverage this kind of multi-fidelity data,[75] with the aim of directly accessing the expensive data at high accuracy by leveraging cheap counterparts at low accuracy. In the present work, a correction learning (CL) scheme is investigated. This method consists in learning the difference between the low- (LDA) and the high-fidelity (HSE) data. Although conceptually simple, this multi-fidelity technique was shown to outperform others when modelling the band gap with MODNet.[61] Since the band gap task has already been addressed in the literature, a MODNet ensemble is chosen to explore the SHG correction in a supervised learning scheme targeting d_{corr}, which is defined as:

$$d_{corr} = d_{LDA} - d_{HSE}, \quad (5)$$

where d_{LDA} is the usual LDA KP coefficient from the main AL dataset and d_{HSE} is the scissor-corrected KP coefficient as introduced above. The feature set is mmf_pgnn, although the inclusion of the following quantities as descriptors is also considered: the LDA gap (E_g^{LDA}), the HSE gap (E_g^{HSE}), the scissor shift (ΔE_g) and d_{LDA}. The feature selection algorithm is the same as that used for the AL and the benchmark of MODNet against the other models (see Section A2, ESI†). The performance of the model is determined *via* a nested 5-fold cross-validation scheme, in which the inner loop implements a hyperparameter optimisation with the native genetic algorithm implemented in the MODNet package.[61]

3 Results

3.1 Conclusion of the active learning procedure

Following the methodology of Section 2.3, almost 20 AL iterations were carried out, two of which consisted of adding materials from Trinquet *et al.*[58] The maximal and minimal numbers of oracle evaluations per iteration were $\sim$280 and $\sim$50, respectively. The performance of the ML model is monitored at each cycle and plots showing the raw metrics are provided in the Section A3 (ESI†). While Fig. A4–A6 (ESI†) correspond to an estimation of the performance from a nested 5-folds cross-validation scheme, the parity plots in Fig. A7–A9 (ESI†) are a better reflection of the reality as they correspond to the selected set of materials at each cycle. Although the curves are not monotonically decreasing, both illustrations show the improvement of the model with the increasing dataset size for all considered metrics. Fig. A10 (ESI†) illustrates the quality of the ML predictions by showing the selected subset of the last fully completed AL iteration in panel (a) and (b). These materials were deemed promising by the selection algorithm based on their $\mu_{d_{KP}}$, $\sigma_{d_{KP}}$, and band gap, which prompted their computation using DFPT. From panel (a), it is clear that the ML predictions are not quantitatively accurate as they sometimes display large discrepancies with the calculated values. However, this level of prediction is still qualitatively good with a high Spearman coefficient (0.89) and is sufficient when performing screening based on relative rankings as emphasised by panel (b). This alternative visualisation compares the predicted and true ranks of each entry and provides a finer insight than the global Spearman coefficient. The correct trend is observed, although it is not perfect. The last panel (c) displays these newly annotated entries in the (E_g, d_{KP}) space on top of the whole dataset at the time. It shows that most of the selected materials are located in the targeted region of the space, despite the exploration regimes of the selection algorithm.

However, one shortcoming of these raw performance checks is the modification of the test sets throughout the AL scheme. To alleviate this issue, a post-processing approach was used to rationalise model performance. The starting training set, T_0, is first divided into 5 folds, $t_{0,j}$. Each of them is then extended by a part of the new DFPT data of each AL iteration:

$$t_{i+1,j} = t_{i,j} \cup f_{i,j}, \quad (6)$$

where $f_{i,j}$ results from a 5-fold splitting of $\mathscr{F}_{DFPT,i}$. Finally, a nested cross-validation scheme is applied on all T_i using the $t_{i,j}$ splitting, which yielded fitted MODNet models, $m_{i,j}$. Each of these sets of models can then be used to perform a cross-validation (without training) of the other T_i. The set of features was restricted to the mmf descriptors. After compiling the results, Fig. 2 and Fig. A11, A12 (ESI†) are obtained. The horizontal axis refers to the index of the models over the AL iterations and the vertical axis indicates the metric. As indicated by the colour, each curve corresponds to

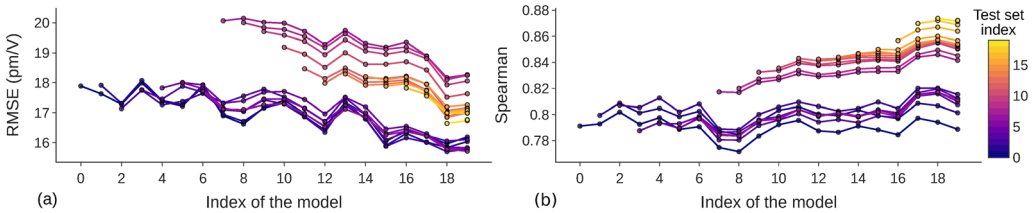

Fig. 2 Evolution of the average RMSE (pm V^{-1}) (a) and Spearman's rank coefficient (b) over the AL process. The index of the model refers to its training set as the AL goes on. Each curve with index i corresponds to the test sets of a 5-folds splitting of the dataset at the ith iteration of the AL procedure such that the same test sets are kept for the whole curve.

a training set T_i. This testing procedure ensures fixed test sets across the AL iterations while avoiding any data leakage. The figures show that all metrics experience an improvement when increasing the training data seen by the models (x-axis). Except for the coefficient of determination, the other metrics present a significant jump when going from the T_6 to the T_7 curves. Both the MAE and RMSE worsen while the Spearman correlation coefficient improves. In the 7th iteration, 239 materials from Trinquet *et al.*[58] were added in the training set, which amounted to 27% of T_6. Moreover, it contained a relatively greater number of high d_{KP} values than the previous additions, which explains the noticeable worsening of performance evidenced by the MAE and RMSE, despite the positive effect on the Spearman coefficient.

The main contribution of the present work to the quest for new NLO inorganic crystals is a new dataset of ~2200 static SHG tensors computed within DFPT, which will be named SHG-25 hereafter. Fig. 3 represents it in the (E_g, d_{KP}) space along with the starting dataset, T_0. From this plot, it is difficult to assess whether SHG-25 properly targets the "Pareto materials" in this space, as intended by the AL procedure.

To appraise this, the Pareto front of T_0 is found and used to fit a function of the form:

$$f_{KP}(E_g) = a \cdot \exp(b \cdot E_g). \quad (7)$$

For each entry x of both T_0 and SHG-25, a normalised distance to this fitted front is then derived as:

$$\Delta d(x) = \frac{f_{KP}(E_g(x)) - d_{KP}(x)}{f_{KP}(E_g(x))}. \quad (8)$$

If it is close to 1, this distance implies that the material is far below the Pareto front, while if it is close to or below 0, the material is in the targeted range of screening. Fig. 4 illustrates the distribution of this proxy target for the two datasets. It shows that SHG-25 contains relatively more compounds close to or above f_{KP} than T_0. The difference is, however, not as striking as it was in Trinquet *et al.*,[58] which can be explained by the low accuracy of the SHG ML model, especially at the start of the data acquisition process. The sampling pool, $\mathscr{P}$, is also more restricted and might not be large enough to effectively push or sample the Pareto front. In addition to this histogram, it is possible to consider the individual data contribution of each AL iteration separately from T_0. To do so, we introduce κ, the fraction of instances with Δd lower than an arbitrary threshold. The latter is set to 0.5 in order to focus on the data closer to the T_0 Pareto fit than to a zero SHG response. In the case of T_0 and SHG-25, κ is equal to 9% and 14%, respectively. When averaged over the first five data contributions of the AL, κ is also 9%, confirming that the first few iterations are almost equivalent to a random selection, as in T_0. However, the last five iterations yield an average κ of 19%, despite materials with high uncertainties being also selected. This demonstrates the performance of our ML model and validates the need to iteratively improve the ML model as the amount of the available training data increases. It is interesting to note that the additions of materials from Trinquet *et al.*[58] display a κ of around 16%, thus confirming the usefulness of targeting compounds with a high refractive index when possible.

The new dataset, SHG-25, is made publicly available on the Materials Cloud Archive[29,30] and on the MPContribs[76] when possible, in the hope that it fosters high-throughput screenings

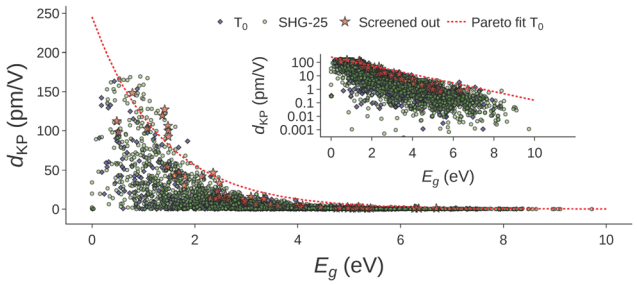

Fig. 3 Representation of the new dataset, SHG-25, and the starting one, T_0, in the (E_g, d_{KP}) space. The red dashed line illustrates the fit of the Pareto front of T_0 while the red stars highlight materials, that are suggested as promising based on several criteria (see text). The band gaps are taken from the source databases (GGA-PBE).

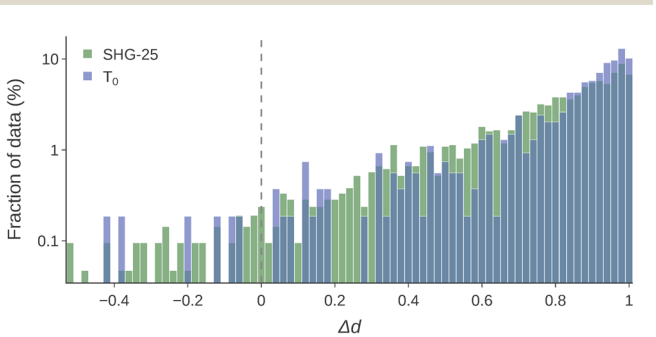

Fig. 4 Percentage of the data binned over the normalised distance from the fitted Pareto front of T_0 as defined in the text.

as well as the development of reliable ML models. Combined with T_0, this new dataset amounts to 2700 entries and is enough to achieve qualitative predictions of the KP coefficient as shown from the above analyses. While high-throughput screenings might already benefit from such accuracy, it is desirable to further improve the performance of cheap ML predictions. Increasing the amount of SHG data is thus of paramount importance.

3.2 Machine-learning the SHG coefficient

In addition to the amount and diversity of training data, the choice of the ML model is another critical factor in the reliability of the d_{KP} predictions. This section presents the results of the ML benchmarks following the methodology introduced in Section 2.4. Table 1 presents the top-level metrics on the largest and most diverse holdout set, sorted by decreasing Spearman's rank correlation coefficient. This performance metric is emphasised as we consider the relative ranking of the predictions to be the most important criterion for screening purposes.

Based on Spearman's rank correlation alone, we find MODNet to be the most performant model ($r_s = 0.87$), also possessing the lowest MAE of 5.76 pm V^{-1} and highest R^2 of 0.70. However, we also find that several models perform competitively with MODNet at this dataset size, both those with increased complexity, namely the co(N)GN series of the graph neural networks (GNNs), and simpler tree-based methods that use the same descriptors as MODNet, namely extra trees and light gradient boosting machines, in agreement with An *et al.*[77] Given both the skewed distribution of d_{KP} values in SHG-25, and the multi-objective nature of our materials design problem (*i.e.*, finding materials on the (E_g, d_{KP}) Pareto front), we also compute enrichment factors (EF) and discovery curves for each model. Using the procedure outlined in Section 3.1, a figure of merit (FOM) for discovery was computed as the distance of a given candidate material from the fitted T_0 Pareto front, Δd.

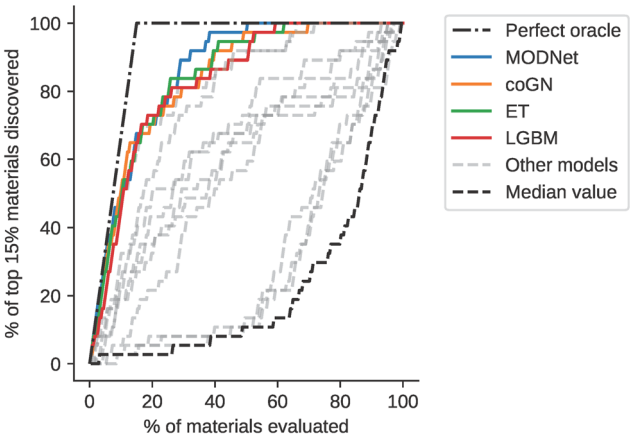

Fig. 5 Discovery curves for the benchmarked models on the top 15% of compounds in the holdout set.

Fig. 5 shows the discovery curves for the top 15% of materials according to the computed FOM, highlighting the performance of the best models. Once again, MODNet, the tree-based methods and the co(N)GN series outcompete all the rest on this metric, achieving normalised EF (15%) values between 0.61 and 0.67, *i.e.*, after evaluating 15% of the dataset following these model's rankings, between 61% and 67% of the top 15% of materials can be recovered. This metric is better suited for capturing model performance specifically when used as a discriminator for potential SHG materials. The clustering around this value perhaps indicates a reasonable maximum enrichment for this holdout set, given the small sample size involved (250 candidates in the holdout set, 37 in the top 15% and thus 12 "missing" from the predictions). MODNet is marginally more efficient at spanning the entire top 15%, requiring around 40% of all materials to be evaluated. Interestingly, even models that perform reasonably well when

Table 1 Performance metrics for the benchmarked models on the SHG-25 dataset for the holdout set, sorted by Spearman's rank correlation coefficient, r_s. In cases where multiple hyperparameter sets or architectures were benchmarked for the same model type, the table presents the model with the best performance. The normalised enrichment factor for the top 15% of materials, EF (15%) is a relevant metric for the application of these models for materials discovery (*e.g.*, continuing the active learning procedure in this study). Standard metrics, mean absolute errors (MAE), root-mean-square errors (RMSE), coefficient of determination (R^2) are also provided for completeness

	MAE (pm V^{-1})	RMSE (pm V^{-1})	r_s	R^2	EF(15%)
MODNet	5.80	15.30	0.87	0.70	0.67
coNGN	6.00	15.50	0.86	0.62	0.61
coGN	6.00	15.10	0.85	0.64	0.64
ET	6.70	15.80	0.85	0.61	0.64
LGBM	6.40	14.70	0.83	0.66	0.64
TensorNet	7.90	16.80	0.79	0.60	0.41
Matten	8.20	20.60	0.79	0.34	0.38
MEGNet	9.30	18.80	0.66	0.44	0.14
Claude Sonnet 3.5	11.60	26.30	0.60	-0.10	0.27
GPT-4o	12.00	27.40	0.52	-0.17	0.30
DARWIN 1.5	13.30	30.00	-0.08	-0.22	0.05

looking at simple metrics like MAE and r_s appear much less effective at this task, with significantly reduced enrichment factors at this threshold.

The threshold of benchmarking against the top 15% of materials is somewhat arbitrary and dataset-dependent. Fig. A2 (ESI†) shows the materials that were selected as the top 15% of this holdout set using the computed FOM. Given the small holdout set size, the choice of threshold is affected by aliasing, however the best-performing models came out on top for all tested thresholds, providing a *post hoc* rationalisation of our choice to use MODNet during the AL procedure.

3.3 Correcting the band gap

Following the selection and computations described in Section 2.5, ~700 pairs of LDA and HSE band gaps as well as the corresponding scissor-corrected SHG tensors are obtained from the ~1000 initially selected compounds and made publicly available along with SHG-25. This new dataset is represented in Fig. 6. As expected, the band gap correction induces a blue-shift of the band gap and decreases the KP coefficient. It can already be seen from this plot that the distribution of d_{HSE} is similar to the non-corrected one. These observations are confirmed by the parity plots in Fig. 7 and Fig. A13 (ESI†). As indicated by the high Spearman's rank correlation coefficients, both figures show that the relative rankings stay the same for both the low and high-fidelity coefficients. This implies that any screening performed at the LDA level is equivalent to screening HSE results. Moreover, the HSE band gaps and the scissor-corrected d_{KP} can both be modelled with a linear regression of their LDA counterpart as a first approximation. Fig. 7 and Fig. A13 (ESI†) illustrate this simple fit by the green dotted line, whose parameters are given in the green box. Given a material with its LDA band gap and its KP coefficient, it is thus possible to approximate its HSE gap and its corresponding KP coefficient at a very low cost.

Unfortunately, these linear regressions present an obvious limitation. For example, any two different materials with two different gap corrections would have the same corrected KP coefficient if their d_{LDA} values are equal. It is thus necessary to go one step further. As described in Section 2.5, a multi-fidelity correction-learning scheme is adopted such that the material-dependent corrections to d_{LDA} are predicted by a MODNet ensemble. Following a nested 5-fold cross-validation, the performance of this approach is investigated by varying the different features sets considered by the ML model. The results are summarised in Table 2 and are fully presented in Table A1 (ESI†). The linear regression introduced above is indicated as a baseline. The scores are derived from $d_{\mathrm{LDA}} - d_{\mathrm{corr}}$ instead of just the correction. This simplifies the interpretation and allows to consider the fraction of predicted corrections with a wrong sign (η) or inducing a negative "HSE" KP coefficient (ζ). These two quantities act as a safeguard against non-physical predictions.

As expected from Fig. 7, the LDA KP coefficient is a necessary feature for the ML model to perform as well as the linear regression, while the band gaps and the scissor are not sufficient. This is not an issue in itself, since d_{LDA} is a prerequisite for using the correction. To further improve MODNet, the band gaps and scissor are separately added as features. As intuition suggests, the scissor results in a significant improvement in the model performance. Further combining all of our custom features only slightly reduces the errors. Unfortunately, the predictions of MODNet are not constrained, as reflected in its η and ζ values of 1% and 3–6%, respectively. In contrast, the linear regression reaches 0% by definition. For this reason, the low values of d_{HSE} are better represented by the linear regression than by MODNet while the higher values benefit from the flexibility of the ML model as illustrated in Fig. A14–A17 (ESI†), although this can be remedied by a simple output rescaling. The significant reduction of the RMSE also supports this interpretation. Thanks to the close relationship between d_{LDA}, d_{HSE} and the custom features, only less than 700 data entries are sufficient to reasonably correct the LDA KP coefficient. This limited dataset size suggests that increasing the data will significantly improve the correction.

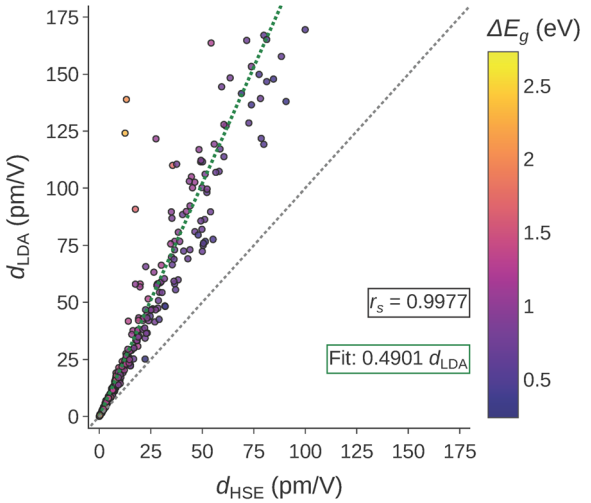

Fig. 7 Parity plot showing the effect of the scissor shift on the KP coefficients. The colour bar indicates the scissor of each compound to go from the LDA to the HSE gap. A linear regression is fitted on those points as shown by the green dotted line and the formula in the box. The Spearman's rank correlation coefficient, r_s, is included as well.

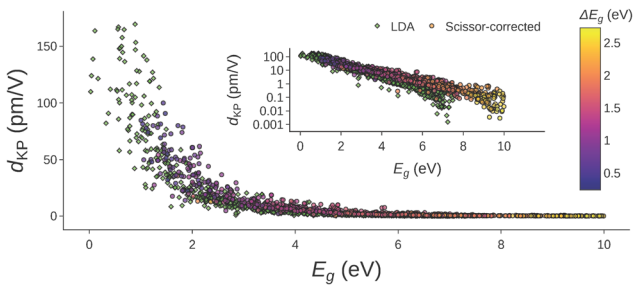

Fig. 6 The ~700 materials subset selected for scissor correction in the (E_g, d_{KP}) space. The inset shows a logarithmic scale for a clearer visualisation. Both the LDA and the scissor-corrected values for the KP coefficients and band gaps are displayed for comparison. The colour bar indicates the scissor of each compound to go from the LDA to the HSE gap.

Table 2 Performance of MODNet on the d_{corr} task when using different set of features under a nested 5-folds cross-validation. The quantities η (%) and ζ (%) correspond to the fraction of d_{corr} with a wrong sign and of negative ($d_{LDA} - d_{corr}$), respectively

	MAE (pm V^{-1})	RMSE (pm V^{-1})	Spearman	R^2	η (%)	ζ (%)
Linear regression	1.6742	4.6544	0.9972	0.9173	0.0000	0.0000
mmf_pgnn	3.5342	9.3651	0.8895	0.6861	0.1481	10.3355
mmf_pgnn $\cup$ d_{LDA}	1.4700	4.3213	0.9837	0.9285	0.7397	5.7614
mmf_pgnn $\cup$ d_{LDA} $\cup$ ΔE_g	1.0984	2.9003	0.9870	0.9659	0.5893	4.1307
mmf_pgnn $\cup$ ΔE_g $\cup$ d_{LDA} $\cup$ E_g^{LDA} $\cup$ E_g^{HSE}	1.0766	2.6977	0.9882	0.9711	1.1776	3.2484

3.4 Screening promising compounds for SHG

As many studies have shown before, the HSE band gaps and the scissor-corrected SHG tensors can successfully be used to screen promising NLO materials with balanced properties.[13,14] Here, this approach is illustrated on our high-fidelity subset of SHG tensors, which contains optimal materials in the (E_g, d_{KP}) space with a gap greater than 1 eV. The screening is based on the following criteria:

- good theoretical stability ($E_{hull} \leq 10$ meV atom^{-1} with respect to the DFT-predicted convex hull of known materials),
- a scissor-corrected KP coefficient (d_{HSE}) greater than 0.33 pm V^{-1},
- a birefringence (Δn_{HSE}) larger than 0.03,
- non-toxic and sustainable elements.

The KP coefficient threshold corresponds to the effective coefficient of the experimental SHG tensor component for the widely used material KH_2PO_4 (KDP), which sets a lower bound for DUV crystals.[78] The birefringence is also restricted by the minimal value for practical application.[79] This condition is challenging to assess since our DFPT calculations only compute the static limit of the electronic contribution to the dielectric tensor. One could argue that the dispersion of the refractive indices is weak below the band gap, thus limiting the difference between the birefringence in the static limit and at a finite frequency. Although Wang *et al.*[14] showed that static birefringence underestimates its counterpart at finite frequencies, the relationship between the two quantities warrants further investigation. To further reduce the selection, compositions with toxic elements (Pb, As, Be, Hg, Cd) or hydrogen were discarded. Moreover, only sustainable elements were retained as characterised by Herfindahl–Hirschman indices (HHIs) lower than 6000 for both production and reserves.‡[80–82]

After applying these criteria, 59 compounds remain, listed in Table A2 (ESI†). As initially desired, the materials selected by these criteria span a broad band gap range, from 1.3 to 9.2 eV, allowing the potentially exemplary materials to be suggested for specific portions of the spectrum relevant to a given application. Our approach is validated by the presence of two well-known nonlinear optical crystals in the list, $Ba(BO_2)_2$ (mp-5730) and LiB_3O_5 (mp-3660). According to the Materials Project,[16] 22 out of the 30 MP entries in the list have already been experimentally observed (*i.e.*, these structures have corresponding entries in the ICSD[22]) and some have already been highlighted as potential NLO materials at the HSE level by Chu *et al.*[13] and Wang *et al.*,[14] as indicated in the table. This highlights the importance of diversifying the original sources of the compounds as well as the ability to periodically reassess the screening with machine-actionable queries of updated databases (*via* OPTIMADE or otherwise). We end this section by briefly highlighting some of the most promising compounds in different areas of the spectrum from Table A2 (ESI†).

In the near-infrared, metastable InP-$P6_3mc$ (wurtzite) is predicted to exhibit a strong NLO response with high birefringence given its small band gap of 1.26 eV; wurtzite InP nanocrystals and nanowires, which can be grown *via* cation exchange, have already attracted interest for optoelectronic applications.[83,84] Chalcopyrite-like $ZnSnP_2$-$P\bar{4}m2$ also falls in this near-IR range, though may struggle to find application given its propensity for disorder.[85]

Several monoclinic and tetragonal phases of metal chalcogenide $Mg(\{In, Ga, Al\}\{Se, Te\}_2)_2$ arise as candidates from this screening, spanning a wide band gap range between 1.46 and 3 eV. These phases have persistent motifs of Mg^{2+}–$\{Te, Se\}^{2-}$ tetrahedra, corner-sharing with $\{In, Ga, Al\}^{3+}$–$\{Te, Se\}^{2-}$ tetrahedra. These phases all have very high predicted d_{KP} values given their band gaps, with birefringences exceeding 0.03. Several of the compositions listed here (and their Mn-based counterparts) have been reported in the literature,[86–88] suggesting that the purely hypothetical compositions (primarily from Alexandria) have a good chance of being synthesised. In this composition space, only $Mg(GaTe_2)_2$ appears in the optical materials literature;[89] interestingly this compound was only recently proposed and isolated and does not yet appear in the ICSD or MP.

The suggested candidates with larger band gaps beyond the visible spectrum in the near- and deep-UV are dominated by borates such as the well-known $Ba(BO_2)_2$ and LiB_3O_5, with more complex quaternary phases still exceeding the predicted d_{KP} threshold set by KDP, such as the carbonates $Na_2Ca_2(CO_3)_3$ and $CaMg_3(CO_3)_4$[90] and fluorooxoborates $Ba_3B_6O_{11}F_2$, $Sr_3B_6O_{11}F_2$ and $Li_2B_6O_9F_2$, the latter of which has been confirmed experimentally as a NLO material with high d_{KP} (90% that of KDP) and large birefringence (0.07).[91]

Given these successes in identifying previously studied NLO compounds, including those not present in experimental databases, we believe that many of the other suggested phases could provide fruitful directions for further study. More work is

‡ Namely Li, B, C, N, O, F, Na, Mg, Al, Si, P, S, Cl, Ca, Zn, Ga, Ge, As, Se, Sr, Cd, In, Sn, Te, I, Ba, Hg, Pb.

required to investigate some of the purely hypothetical compounds in this list to assess not only their potential NLO properties, but also their synthesisability.

4 Conclusions and outlook

Despite a large and active community, the field of nonlinear optical materials is still looking for appropriate compounds in specific electromagnetic ranges, such as the deep UV and the mid- and far-infrared, that could be used in industrial applications. Today, this search can be driven by computations in order to accelerate the discovery of promising compounds.[8] In order to navigate the rapidly growing design space offered by curated databases of hypothetical compounds, it is imperative to use fast screening methods to avoid wasting computational resources on suboptimal materials. One solution is to train cheap machine learning models on the target property to efficiently guide the allocation of DFT resources. However, this approach necessitates a large enough pre-existing dataset of the target property to attain a reasonable predictive power. Since the field of NLO materials lacks datasets, the present work adopted an active learning framework to acquire new static SHG tensors. By leveraging a relatively small existing dataset, this procedure resulted in ∼2200 newly computed SHG tensors, which is made openly available on the materials could archive[30] and is itself accessible *via* an OPTIMADE API.[27] The ML proxy allowed us to bias the data acquisition towards compounds exhibiting high SHG coefficients given their band gap. Thanks to this new dataset, we were able to test a variety of ML models on this SHG task and its relationship with higher-fidelity data was also investigated. The tools used throughout this work enable periodic reassessment of the decentralised design space with minimal modifications to the code. This has already begun, as shown in Fig. 8, where the GNoME dataset[24] (∼10 000 relevant entries) and new entries in Alexandria (∼20 000 relevant entries) have been screened using our latest model.

Although the effective KP coefficient can be qualitatively predicted with the present ML model, it is of interest to the community to improve its performance. We believe that the first step in achieving this is to increase the amount of training data. Thanks to the OPTIMADE API, we plan on continually extending our SHG dataset by querying unexplored databases providing either experimentally verified compounds or hypothetical compounds with the proper thermodynamic information. If the source does not provide band gaps, then one of the many ML models in the literature can be used to approximate it. In parallel to the screening of existing data, we could try to generate our own pool of hypothetical compounds. On the one hand, more targeted searches for hypothetical stable materials can make use of an evolutionary algorithm before being filtered on predicted SHG coefficients.[77,92] On the other hand, inverse design *via* constrained generation might quickly offer suggestions of promising compositions and/or structures.[93,94]

In addition to the data-driven identification of suitable compounds, the present dataset could be leveraged to derive physical insights and better understand the characteristics behind a good NLO material. Whilst a close investigation of the promising candidates is not within the scope of this work, we invite the community to consider these compounds, pending more detailed calculations of their suitability in future work.

In the search for new or yet-to-be-investigated functional materials, it is essential to consider the main requirements for practical applications. In the realm of bulk SHG crystals, critical considerations include achieving robust SHG conversion, the capability for angular phase-matching (APM), and a high laser damage threshold (LDT). Since the present study is based on first-principles computations, we emphasised the bi-objective optimisation of the effective powder SHG coefficient (d_{KP}) and the band gap, in an attempt to maximise the conversion efficiency and LDT. Given the scale of our results, this approach constitutes a considerable step in the data-driven search for NLO materials. With this dataset now established, we hope to directly incorporate additional proxies for real-world performance, such as the lattice thermal conductivity and birefringence,[13,95–98]

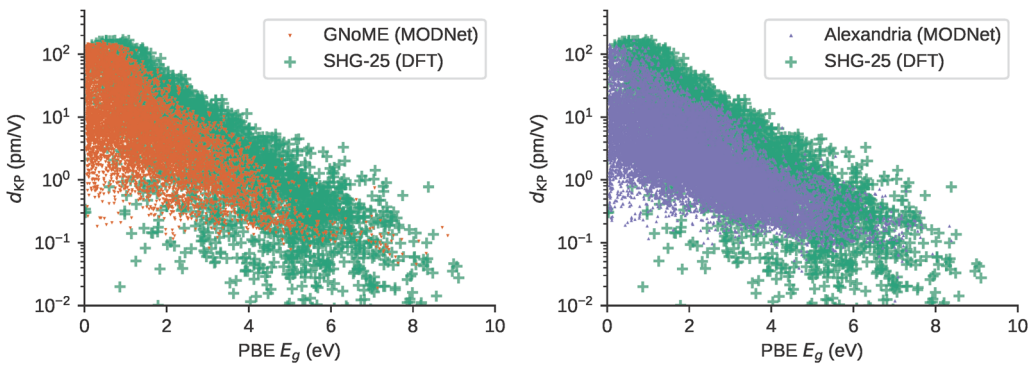

Fig. 8 MODNet-predicted d_{KP} values for hypothetical structures added to GNoME (left) and Alexandria (right) since the conclusion of the active learning study, plotted against the database-reported band gaps computed at the PBE level, alongside the DFT-computed SHG values in SHG-25 (green). The structures considered were limited to those that: (i) are near the predicted convex hull reported in the database (≤0.05 eV per atom), (ii) have PBE band gaps greater than 0.05 eV, (iii) are non-centrosymmetric, (iv) do not contain lanthanides or actinides, and (v) have compositions that are not present in the computed SHG set. This left 9657 structures from GNoME and 22 438 from Alexandria.

in our future work. These factors will reinforce the focus on materials with high LDT and APM capability, respectively, thereby refining the data acquisition process, expanding our dataset, and broadening the applicability of these findings.

Author contributions

V. T.: conceptualisation, methodology, software, validation, investigation, data curation, writing – original draft, writing – review and editing, visualisation. M. L. E.: conceptualisation, methodology, software, validation, investigation, data curation, writing – original draft, writing – review and editing, visualisation. G.-M. R.: conceptualisation, resources, writing – review and editing, supervision, funding acquisition.

Conflicts of interest

G.-M. R. is a shareholder and Chief Innovation Officer of Matgenix SRL.

Data availability

The crystal structures, DFT band gaps, and DFPT SHG tensors of both SHG-25 and its scissor-corrected subset have been deposited to the Materials Cloud Archive (**https://doi.org/10.24435/materialscloud:wk-qm**), with a corresponding OPTIMADE API available at **https://optimade.materialscloud.org/archive/materialscloud:wk-qm**. The entries from the MP are also accessible *via* an MPContribs[76] record (**https://contribs.materialsproject.org/projects/shg**). The code repository **https://github.com/modl-uclouvain/shg-ml-benchmarks** (archived at **https://doi.org/10.5281/zenodo.15691912**) contains the ML benchmarking code, results, and the scripts used for screening the latest OPTIMADE structures. It also contains the scripts used for the AL procedure. The code is released under the permissive MIT license. All the links and data can be found on our webpage (**https://nlo.modl-uclouvain.org/**).

Acknowledgements

Computational resources were provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI) funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11 and by the Walloon Region. The present research benefited from computational resources made available on Lucia, the Tier-1 supercomputer of the Walloon Region, infrastructure funded by the Walloon Region under grant agreement no. 1910247. V. T. acknowledges support from the FRS-FNRS through a FRIA Grant. M. L. E. thanks the BEWARE scheme of the Wallonia-Brussels Federation for funding under the European Commission's Marie Curie-Skłodowska Action (COFUND 847587).

Notes and references

1 R. W. Boyd, *Nonlinear Optics*, Academic Press, San Diego, 1992.
2 L. Kang and Z. Lin, Deep-ultraviolet nonlinear optical crystals: concept development and materials discovery, *Light: Sci. Appl.*, 2022, **11**(201), 1–12, DOI: **10.1038/s41377-022-00899-1**, ISSN 2047-7538.
3 L. G. Legres, C. Chamot, M. Varna and A. Janin, The Laser Technology:New Trends in Biology and Medicine, *J. Mod. Phys.*, 2014, **5**(5), 267–279, DOI: **10.4236/jmp.2014.55037**.
4 L. Jia, J. Wu, Y. Zhang, Y. Qu, B. Jia and D. J. Moss, Third-Order Optical Nonlinearities of 2D Materials at Telecommunications Wavelengths, *Micromachines*, 2023, **14**(2), 307, DOI: **10.3390/mi14020307**, ISSN 2072-666X.
5 A. Dutt, A. Mohanty, A. L. Gaeta and M. Lipson, Nonlinear and quantum photonics using integrated optical materials, *Nat. Rev. Mater.*, 2024, **9**, 321–346, DOI: **10.1038/s41578-024-00668-z**, ISSN 2058-8437.
6 J. M. Rondinelli and E. Kioupakis, Predicting and Designing Optical Properties of Inorganic Materials, *Annu. Rev. Mater. Res.*, 2015, **45**(1), 491–518, DOI: **10.1146/annurev-matsci-070214-021150**, ISSN 1531-7331, 1545-4118.
7 H. Z. Aslam, J. T. Doane, M. T. Yeung and G. Akopov, Advances in Solid-State Nonlinear Optical Materials: From Fundamentals to Applications, *ACS Appl. Opt. Mater.*, 2023, **1**(12), 1898–1921, DOI: **10.1021/acsaom.3c00352**.
8 H. Wang, M. Mutailipu, Z. Yang, S. Pan and J. Li, Computer-Aided Development of New Nonlinear Optical Materials, *Angew. Chem., Int. Ed.*, 2025, **64**(6), e202420526, DOI: **10.1002/anie.202420526**, ISSN 1433-7851.
9 X. Jiang, L. Kang, S. Luo, P. Gong, M.-H. Lee and Z. Lin, Development of nonlinear optical materials promoted by density functional theory simulations, *Int. J. Mod. Phys. B*, 2014, **28**(27), 1430018, DOI: **10.1142/S0217979214300187**, ISSN 0217-9792.
10 H. Lin, W.-B. Wei, H. Chen, X.-T. Wu and Q.-L. Zhu, Rational design of infrared nonlinear optical chalcogenides by chemical substitution, *Coord. Chem. Rev.*, 2020, **406**, 213150, DOI: **10.1016/j.ccr.2019.213150**, ISSN 0010-8545.
11 X. Dong, L. Huang and G. Zou, Rational Design and Controlled Synthesis of High-Performance Inorganic Short-Wave UV Nonlinear Optical Materials, *Acc. Chem. Res.*, 2025, **58**(1), 150–162, DOI: **10.1021/acs.accounts.4c00704**, ISSN 0001-4842.
12 B. Zhang, X. Zhang, J. Yu, Y. Wang, K. Wu and M.-H. Lee, First-Principles High-Throughput Screening Pipeline for Nonlinear Optical Materials: Application to Borates, *Chem. Mater.*, 2020, **32**(15), 6772–6779, DOI: **10.1021/acs.chemmater.0c02583**, ISSN 0897-4756.
13 D. Chu, Y. Huang, C. Xie, E. Tikhonov, I. Kruglov, G. Li, S. Pan and Z. Yang, Unbiased Screening of Novel Infrared Nonlinear Optical Materials with High Thermal Conductivity:Long-neglected Nitrides and Popular Chalcogenides, *Angew. Chem., Int. Ed.*, 2023, **62**(16), e202300581, DOI: **10.1002/anie.202300581**, ISSN 1433-7851.

14 J. Wang, M. Ye, X. Guo, Y. Li, N. Zou, H. Li, Z. Zhang, S. Zhao, Z. Xu, H. Chen, D. Wu, T. Bao, Y. Xu and W. Duan, Unbiased screening of deep-ultraviolet and mid-infrared nonlinear optical crystals: Long-neglected covalent and mixed-cation motifs, *Phys. Rev. Mater.*, 2024, **8**(8), 085202, DOI: **10.1103/PhysRevMaterials.8.085202**.

15 Y. Alkabakibi, D. D. Barma, D. V. Rybkovskiy, A. Tudi, C. Xie and A. R. Oganov, Computational Identification of Four Promising Nonlinear Optical Materials for Near and Middle Ultraviolet Operation, *JETP Lett.*, 2025, 1–6, DOI: **10.1134/S0021364024605074**, ISSN 1090-6487.

16 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. Davidson Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, *APL Mater.*, 2013, **1**(1), 011002, DOI: **10.1063/1.4812323**.

17 J. Schmidt, L. Pettersson, C. Verdozzi, S. Botti and M. A. L. Marques, Crystal graph attention networks for the prediction of stable materials, *Sci. Adv.*, 2021, **7**(49), eabi7948, DOI: **10.1126/sciadv.abi7948**.

18 J. Yu, B. Zhang, X. Zhang, Y. Wang, K. Wu and M.-H. Lee, Finding Optimal Mid-Infrared Nonlinear Optical Materials in Germanates by First-Principles High-Throughput Screening and Experimental Verification, *ACS Appl. Mater. Interfaces*, 2020, **12**(40), 45023–45035, DOI: **10.1021/acsami.0c15728**, ISSN 1944-8244.

19 R. Wang, F. Liang and Z. Lin, Data-driven prediction of diamond-like infrared nonlinear optical crystals with targeting performances, *Sci. Rep.*, 2020, **10**(3486), 1–8, DOI: **10.1038/s41598-020-60410-x**, ISSN 2045-2322.

20 C. Xie, E. Tikhonov, D. Chu, M. Wu, I. Kruglov, S. Pan and Z. Yang, A prediction-driven database to enable rapid discovery of nonlinear optical materials, *Sci. China Mater.*, 2023, **66**(11), 4473–4479, DOI: **10.1007/s40843-023-2592-x**, ISSN 2199-4501.

21 G. Bergerhoff, R. Hundt, R. Sievers and I. D. Brown, The inorganic crystal structure data base, *J. Chem. Inf. Model.*, 1983, **23**(2), 66–69, DOI: **10.1021/ci00038a003**.

22 D. Zagorac, H. Müller, S. Ruehl, J. Zagorac and S. Rehme, Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features, *J. Appl. Crystallogr.*, 2019, **52**(5), 918–925, DOI: **10.1107/S160057671900997X**.

23 S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl and C. Wolverton, The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies, *npj Comput. Mater.*, 2015, **1**(1), 1–15, DOI: **10.1038/npjcompumats.2015.10**.

24 A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon and E. Dogus Cubuk, Scaling deep learning for materials discovery, *Nature*, 2023, **624**(7990), 80–85, DOI: **10.1038/s41586-023-06735-9**, ISSN 1476-4687.

25 C. Zeni, R. Pinsler, D. Zügner, A. Fowler, M. Horton, X. Fu, S. Shysheya, J. Crabbé, L. Sun, J. Smith, B. Nguyen, H. Schulz, S. Lewis, C.-W. Huang, Z. Lu, Y. Zhou, H. Yang, H. Hao, J. Li, R. Tomioka and T. Xie, *MatterGen: A generative model for inorganic materials design*, *arXiv*, 2024, preprint, arXiv:2312.03687, DOI: **10.48550/arXiv.2312.03687**.

26 A. K. Cheetham and R. Seshadri, Artificial Intelligence Driving Materials Discovery? Perspective on the Article: Scaling Deep Learning for Materials Discovery, *Chem. Mater.*, 2024, **36**(8), 3490–3495, DOI: **10.1021/acs.chemmater.4c00643**.

27 C. W. Andersen, R. Armiento, E. Blokhin, G. J. Conduit, S. Dwaraknath, M. L. Evans, A. Fekete, A. Gopakumar, S. Gražulis, A. Merkys, F. Mohamed, C. Oses, G. Pizzi, G.-M. Rignanese, M. Scheidgen, L. Talirz, C. Toher, D. Winston, R. Aversa, K. Choudhary, P. Colinet, S. Curtarolo, D. D. Stefano, C. Draxl, S. Er, M. Esters, M. Fornari, M. Giantomassi, M. Govoni, G. Hautier, V. Hegde, M. K. Horton, P. Huck, G. Huhs, J. Hummelshøj, A. Kariryaa, B. Kozinsky, S. Kumbhar, M. Liu, N. Marzari, A. J. Morris, A. A. Mostofi, K. A. Persson, G. Petretto, T. Purcell, F. Ricci, F. Rose, M. Scheffler, D. Speckhard, M. Uhrin, A. Vaitkus, P. Villars, D. Waroquiers, C. Wolverton, M. Wu and X. Yang, OPTIMADE, an API for exchanging materials data, *Sci. Data*, 2021, **8**(1), 217, DOI: **10.1038/s41597-021-00974-z**, ISSN 2052-4463.

28 M. Evans, J. Bergsma, A. Merkys, C. Andersen, O. B. Andersson, D. Beltrán, E. Blokhin, T. M. Boland, R. Castañeda Balderas, K. Choudhary, A. Díaz Díaz, R. Domínguez García, H. Eckert, K. Eimre, M. Elena Fuentes-Montero, A. M. Krajewski, J. Jørgen Mortensen, J. Manuel Nápoles-Duarte, J. Pietryga, J. Qi, F. de Jesús Trejo Carrillo, A. Vaitkus, J. Yu, A. Zettel, P. Baptista de Castro, J. Martin Carlsson, T. F. T. Cerqueira, S. Divilov, H. Hajiyani, F. Hanke, K. Jose, C. Oses, J. Riebesell, J. Schmidt, D. Winston, C. Xie, X. Yang, S. Bonella, S. Botti, S. Curtarolo, C. Draxl, L. E. E. Fuentes-Cobas, A. Hospital, Z.-K. Liu, L. Marques Miguel A, N. Marzari, A. James Morris, S. Ping Ong, M. Orozco, K. Persson, K. Sommer Thygesen, C. M. Wolverton, M. Scheidgen, C. Toher, G. Conduit, G. Pizzi, S. Grazulis, G.-M. Rignanese and R. Armiento, Developments and applications of the OPTIMADE API for materials discovery, design, and data exchange, *Digital Discovery*, 2024, **3**(8), 1509–1533, DOI: **10.1039/D4DD00039K**.

29 L. Talirz, S. Kumbhar, E. Passaro, A. V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S. P. Huber, S. Zoupanos, C. S. Adorf, C. Welzel Andersen, O. Schütt, C. A. Pignedoli, D. Passerone, J. VandeVondele, T. C. Schulthess, B. Smit, G. Pizzi and N. Marzari, Materials Cloud, a platform for open computational science, *Sci. Data*, 2020, **7**(1), 299, DOI: **10.1038/s41597-020-00637-5**, ISSN 2052-4463.

30 V. Trinquet, M. L. Evans and G.-M. Rignanese, *Research data supporting "Accelerating the discovery of highperformance nonlinear optical materials using active learning and high-throughput screening"*, 2025, DOI: **10.24435/materialscloud:wk-qm**.

31 F. Zernike and J. E. Midwinter, *Applied Nonlinear Optics*, Wiley, New York, 1973.

32 G. New, *Introduction to Nonlinear Optics*, Cambridge University Press, New York, 2011.
33 S. K. Kurtz and T. T. Perry, A Powder Technique for the Evaluation of Nonlinear Optical Materials, *J. Appl. Phys.*, 1968, **39**(8), 3798–3813, DOI: 10.1063/1.1656857, ISSN 0021-8979.
34 X. Gonze, F. Jollet, F. Abreu Araujo, D. Adams, B. Amadon, T. Applencourt, C. Audouze, J.-M. Beuken, J. Bieder, A. Bokhanchuk, E. Bousquet, F. Bruneval, D. Caliste, M. Côté, F. Dahm, F. Da Pieve, M. Delaveau, M. Di Gennaro, B. Dorado, C. Espejo, G. Geneste, L. Genovese, A. Gerossier, M. Giantomassi, Y. Gillet, D. R. Hamann, L. He, G. Jomard, J. Laflamme Janssen, S. Le Roux, A. Levitt, A. Lherbier, F. Liu, I. Lukačević, A. Martin, C. Martins, M. J. T. Oliveira, S. Poncé, Y. Pouillon, T. Rangel, G.-M. Rignanese, A. H. Romero, B. Rousseau, O. Rubel, A. A. Shukri, M. Stankovski, M. Torrent, M. J. Van Setten, B. Van Troeye, M. J. Verstraete, D. Waroquiers, J. Wiktor, B. Xu, A. Zhou and J. W. Zwanziger, Recent developments in the ABINIT software package, *Comput. Phys. Commun.*, 2016, **205**, 106–131, DOI: 10.1016/j.cpc.2016.04.003, ISSN 0010-4655.
35 X. Gonze, B. Amadon, G. Antonius, F. Arnardi, L. Baguet, J.-M. Beuken, J. Bieder, F. Bottin, J. Bouchet, E. Bousquet, N. Brouwer, F. Bruneval, G. Brunin, T. Cavignac, J.-B. Charraud, W. Chen, M. Côté, S. Cottenier, J. Denier, G. Geneste, P. Ghosez, M. Giantomassi, Y. Gillet, O. Gingras, D. R. Hamann, G. Hautier, X. He, N. Helbig, N. Holzwarth, Y. Jia, F. Jollet, W. Lafargue-Dit-Hauret, K. Lejaeghere, M. A. L. Marques, A. Martin, C. Martins, H. P. C. Miranda, F. Naccarato, K. Persson, G. Petretto, V. Planes, Y. Pouillon, S. Prokhorenko, F. Ricci, G.-M. Rignanese, A. H. Romero, M. Marcus Schmitt, M. Torrent, M. J. van Setten, B. Van Troeye, M. J. Verstraete, G. Zérah and J. W. Zwanziger, The Abinit project:Impact, environment and recent developments, *Comput. Phys. Commun.*, 2020, **248**, 107042, DOI: 10.1016/j.cpc.2019.107042, ISSN 0010-4655.
36 A. H. Romero, D. C. Allan, B. Amadon, G. Antonius, T. Applencourt, L. Baguet, J. Bieder, F. Bottin, J. Bouchet, E. Bousquet, F. Bruneval, G. Brunin, D. Caliste, M. Côté, J. Denier, C. Dreyer, P. Ghosez, M. Giantomassi, Y. Gillet, O. Gingras, D. R. Hamann, G. Hautier, F. Jollet, G. Jomard, A. Martin, H. P. C. Miranda, F. Naccarato, G. Petretto, N. A. Pike, V. Planes, S. Prokhorenko, T. Rangel, F. Ricci, G.-M. Rignanese, M. Royo, M. Stengel, M. Torrent, M. J. van Setten, B. Van Troeye, M. J. Verstraete, J. Wiktor, J. W. Zwanziger and X. Gonze, ABINIT: Overview and focus on selected capabilities, *J. Chem. Phys.*, 2020, **152**(12), 124102, DOI: 10.1063/1.5144261, ISSN 0021-9606.
37 X. Gonze, Perturbation expansion of variational principles at arbitrary order, *Phys. Rev. A: At., Mol., Opt. Phys.*, 1995, **52**(2), 1086–1095, DOI: 10.1103/PhysRevA.52.1086, ISSN 2469-9934.
38 M. Veithen, X. Gonze and P. Ghosez, Nonlinear optical susceptibilities, Raman efficiencies, and electrooptic tensors from first-principles density functional perturbation theory, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2005, **71**(12), 125107, DOI: 10.1103/PhysRevB.71.125107, ISSN 2469-9969.
39 M. J. van Setten, M. Giantomassi, E. Bousquet, M. J. Verstraete, D. R. Hamann, X. Gonze and G.-M. Rignanese, The PseudoDojo: Training and grading a 85 element optimized norm-conserving pseudopotential table, *Comput. Phys. Commun.*, 2018, **226**, 39–54, DOI: 10.1016/j.cpc.2018.01.012, ISSN 0010-4655.
40 J. P. Perdew and Y. Wang, Accurate and simple analytic representation of the electron-gas correlation energy, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1992, **45**(23), 13244–13249, DOI: 10.1103/PhysRevB.45.13244.
41 A. M. Ganose, H. Sahasrabuddhe, M. Asta, K. Beck, T. Biswas, A. Bonkowski, J. Bustamante, X. Chen, Y. Chiang, D. C. Chrzan, J. Clary, O. A. Cohen, C. Ertural, M. C. Gallant, J. George, S. Gerits, R. E. A. Goodall, R. D. Guha, G. Hautier, M. Horton, T. J. Inizan, A. D. Kaplan, R. S. Kingsbury, M. C. Kuner, B. Li, X. Linn, M. J. McDermott, R. Srinivaas Mohanakrishnan, A. N. Naik, J. B. Neaton, S. M. Parmar, K. A. Persson, G. Petretto, T. A. R. Purcell, F. Ricci, B. Rich, J. Riebesell, G.-M. Rignanese, A. S. Rosen, M. Scheffler, J. Schmidt, J.-X. Shen, A. Sobolev, R. Sundararaman, C. Tezak, V. Trinquet, J. B. Varley, D. Vigil-Fowler, D. Wang, D. Waroquiers, M. Wen, H. Yang, H. Zheng, J. Zheng, Z. Zhu and A. Jain, Atomate2: modular workflows for materials science, *Digital Discovery*, 2025, **4**(7), 1944–1973, DOI: 10.1039/D5DD00019J, ISSN 2635-098X.
42 A. S. Rosen, M. Gallant, J. George, J. Riebesell, H. Sahasrabuddhe, J.-X. Shen, M. Wen, M. L. Evans, G. Petretto, D. Waroquiers, G.-M. Rignanese, K. A. Persson, A. Jain and A. M. Ganose, Jobflow: Computational Workflows Made Simple, *J. Open Source Software*, 2024, **9**(93), 5995, DOI: 10.21105/joss.05995, ISSN 2475-9066.
43 V. Trinquet, F. Naccarato, G. Brunin, G. Petretto, L. Wirtz, G. Hautier and G.-M. Rignanese, Second-harmonic generation tensors from high-throughput density-functional perturbation theory, *Sci. Data*, 2024, **11**(757), 1–10, DOI: 10.1038/s41597-024-03590-9, ISSN 2052-4463.
44 A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G.-M. Rignanese, G. Hautier, D. Gunter and K. A. Persson, FireWorks: a dynamic workflow system designed for high-throughput applications, *Concurrency Comput.: Pract. Exper.*, 2015, **27**(17), 5037–5059, DOI: 10.1002/cpe.3505, ISSN 1532-0626.
45 IEEE Standard on Piezoelectricity. ANSI/IEEE Std 176-1987, p. 0_1, 1988, DOI: 10.1109/IEEESTD.1988.79638.
46 D. A. Roberts, Simplified characterization of uniaxial and biaxial nonlinear optical crystals: a plea for standardization of nomenclature and conventions, *IEEE J. Quantum Electron.*, 1992, **28**(10), 2057–2074, DOI: 10.1109/3.159516.
47 J. P. Perdew, K. Burke and M. Ernzerhof, Generalized Gradient Approximation Made Simple, *Phys. Rev. Lett.*, 1996, **77**(18), 3865–3868, DOI: 10.1103/PhysRevLett.77.3865, ISSN 1079-7114.
48 J. P. Perdew, K. Burke and M. Ernzerhof, Generalized Gradient Approximation Made Simple [Phys. Rev. Lett. 77,

3865 (1996)], *Phys. Rev. Lett.*, 1997, **78**(7), 1396, DOI: 10.1103/PhysRevLett.78.1396, ISSN 1079-7114.

49 J. Heyd, G. E. Scuseria and M. Ernzerhof, Hybrid functionals based on a screened coulomb potential, *J. Chem. Phys.*, 2003, **118**(18), 8207–8215, DOI: 10.1063/1.1564060, ISSN 1089-7690.

50 J. Heyd, G. E. Scuseria and M. Ernzerhof, Erratum: "Hybrid functionals based on a screened coulomb potential" [J. Chem. Phys. 118, 8207 (2003)], *J. Chem. Phys.*, 2006, **124**(21), 219906, DOI: 10.1063/1.2204597, ISSN 1089-7690.

51 P. E. Blöchl, Projector augmented-wave method, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1994, **50**(24), 17953–17979, DOI: 10.1103/PhysRevB.50.17953, ISSN 2469-9969.

52 G. Kresse and J. Hafner, Norm-conserving and ultrasoft pseudopotentials for first-row and transition elements, *J. Phys.: Condens. Matter*, 1994, **6**(40), 8245, DOI: 10.1088/0953-8984/6/40/015, ISSN 0953-8984.

53 G. Kresse and D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1999, **59**(3), 1758–1775, DOI: 10.1103/PhysRevB.59.1758, ISSN 2469-9969.

54 S. L. Dudarev, G. A. Botton, S. Y. Savrasov, C. J. Humphreys and A. P. Sutton, Electron-energy-loss spectra and the structural stability of nickel oxide: An LSDA + U study, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1998, **57**(3), 1505–1509, DOI: 10.1103/PhysRevB.57.1505, ISSN 2469-9969.

55 A. Jain, G. Hautier, C. J. Moore, S. Ping Ong, C. C. Fischer, T. Mueller, K. A. Persson and G. Ceder, A high-throughput infrastructure for density functional theory calculations, *Comput. Mater. Sci.*, 2011, **50**(8), 2295–2310, DOI: 10.1016/j.commatsci.2011.02.023, ISSN 0927-0256.

56 E. Luppi, H. Hübener and V. Véniard, Ab initio second-order nonlinear optics in solids: Second-harmonic generation spectroscopy from time-dependent density-functional theory, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2010, **82**(23), 235201, DOI: 10.1103/PhysRevB.82.235201.

57 C. Attaccalite and M. Grüning, Nonlinear optics from an ab initio approach by means of the dynamical Berry phase: Application to second- and third-harmonic generation in semiconductors, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2013, **88**(23), 235113, DOI: 10.1103/PhysRevB.88.235113.

58 V. Trinquet, M. L. Evans, C. J. Hargreaves, P.-P. De Breuck and G.-M. Rignanese, Optical materials discovery and design with federated databases and machine learning, *Faraday Discuss.*, 2025, **256**, 459–482, DOI: 10.1039/D4FD00092G, ISSN 1359-6640.

59 P.-P. De Breuck, G. Hautier and G.-M. Rignanese, Materials property prediction for limited datasets enabled by feature selection and joint learning with MODNet, *npj Comput. Mater.*, 2021, **7**(1), 83, DOI: 10.1038/s41524-021-00552-2, ISSN 2057-3960.

60 P.-P. De Breuck, M. L. Evans and G.-M. Rignanese, Robust model benchmarking and bias-imbalance in data-driven materials science: A case study on MODNet, *J. Phys.: Condens. Matter*, 2021, **33**(40), 404002, DOI: 10.1088/1361-648X/ac1280, ISSN 0953-8984.

61 P.-P. De Breuck, G. Heymans and G.-M. Rignanese, Accurate experimental band gap predictions with multifidelity correction learning, *J. Mater. Inf.*, 2022, **2**, 10, DOI: 10.20517/jmi.2022.13, ISSN 2770-372X.

62 L. Ward, A. Dunn, A. Faghaninia, N. E. R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. Jeffrey Snyder, I. Foster and A. Jain, Matminer: An open source toolkit for materials data mining, *Comput. Mater. Sci.*, 2018, **152**, 60–69, DOI: 10.1016/j.commatsci.2018.05.018, ISSN 0927-0256.

63 R. Gouvêa, *et al.*, *In preparation*, https://github.com/rogeriog/pGNN.

64 R. C. Miller, Optical second harmonic generation in piezoelectric crystals, *Appl. Phys. Lett.*, 1964, **5**(1), 17–19, DOI: 10.1063/1.1754022, ISSN 0003-6951.

65 J. Schmidt, N. Hoffmann, H.-C. Wang, P. Borlido, P. J. M. A. Carriço, T. F. T. Cerqueira, S. Botti and M. A. L. Marques, Machine-Learning-Assisted Determination of the Global Zero-Temperature Phase Diagram of Materials, *Adv. Mater.*, 2023, **35**(22), 2210788, DOI: 10.1002/adma.202210788, ISSN 1521-4095.

66 S. Ping Ong, W. Davidson Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis, *Comput. Mater. Sci.*, 2013, **68**, 314–319, DOI: 10.1016/j.commatsci.2012.10.028, ISSN 0927-0256.

67 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

68 P. Geurts, D. Ernst and L. Wehenkel, Extremely randomized trees, *Mach. Learn.*, 2006, **63**(1), 3–42, DOI: 10.1007/s10994-006-6226-1, ISSN 1573-0565.

69 G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye and T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in *Advances in Neural Information Processing Systems*, ed. I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, Curran Associates, Inc., 2017, vol. 30, https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

70 R. Ruff, P. Reiser, J. Stühmer and P. Friederich, Connectivity optimized nested line graph networks for crystal structures, *Digital Discovery*, 2024, **3**(3), 594–601, DOI: 10.1039/D4DD00018H, ISSN 2635-098X.

71 C. Chen, W. Ye, Y. Zuo, C. Zheng and S. Ping Ong, Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals, *Chem. Mater.*, 2019, **31**(9), 3564–3572, DOI: 10.1021/acs.chemmater.9b01294.

72 G. Simeon and G. de Fabritiis, TensorNet: Cartesian Tensor Representations for Efficient Learning of Molecular Potentials, *arXiv*, 2023, preprint, arXiv:2306.06482, DOI: 10.48550/arXiv.2306.06482.

73 M. Wen, M. K. Horton, J. M. Munro, P. Huck and K. A. Persson, An equivariant graph neural network for the elasticity tensors of all seven crystal systems, *Digital Discovery*, 2024, **3**(5), 869–882, DOI: **10.1039/D3DD00233K**, ISSN 2635-098X.

74 T. Xie, Y. Wan, Y. Liu, Y. Zeng, S. Wang, W. Zhang, C. Grazian, C. Kit, W. Ouyang, D. Zhou and B. Hoex, DARWIN 1.5: Large Language Models as Materials Science Adapted Learners, *arXiv*, 2024, preprint, arXiv:2412.11970, DOI: **10.48550/arXiv.2412.11970**.

75 J. Kim, J. Kim, J. Kim, J. Lee, Y. Park, Y. Kang and S. Han, Data-Efficient Multifidelity Training for High-Fidelity Machine Learning Interatomic Potentials, *J. Am. Chem. Soc.*, 2025, **147**(1), 1042–1054, DOI: **10.1021/jacs.4c14455**, ISSN 0002-7863.

76 P. Huck, A. Jain, D. Gunter, D. Winston and K. Persson, A Community Contribution Framework for Sharing Materials Data with Materials Project, In 2015 IEEE 11th International Conference on e-Science, 2015, pp. 535–541, DOI: **10.1109/eScience.2015.75**.

77 R. An, H. Wang, C. Xie, M. Wu, D. Chu, W. Jin, J. Li, S. Pan and Z. Yang, New Ways to Discover Novel Nonlinear Optical Materials: Scaling Machine Learning with Chemical Descriptors Information, *Small*, 2025, **21**(11), 2500540, DOI: **10.1002/smll.202500540**, ISSN 1613-6810.

78 V. G. Dmitriev, G. G. Gurzadyan and D. N. Nikogosyan, *Handbook of Nonlinear Optical Crystals*, Springer, Berlin, Germany, 1999, ISBN 978-3-540-46793-9. **https://link.springer.com/book/10.1007/978-3-540-46793-9**.

79 E. Luppi and V. Véniard, A review of recent theoretical studies in nonlinear crystals: towards the design of new materials, *Semicond. Sci. Technol.*, 2016, **31**(12), 123002, DOI: **10.1088/0268-1242/31/12/123002**, ISSN 0268-1242.

80 O. C. Herfindahl, *Concentration In The Steel Industry*, PhD thesis, Columbia University, 1950.

81 M. W. Gaultois, T. D. Sparks, C. K. H. Borg, R. Seshadri, W. D. Bonificio and D. R. Clarke, Data-Driven Review of Thermoelectric Materials: Performance and Resource Considerations, *Chem. Mater.*, 2013, **25**(15), 2911–2920, DOI: **10.1021/cm400893e**.

82 J. Kim, D. H. Mok, H. Kim and S. Back, Accelerating the Search for New Solid Electrolytes: Exploring Vast Chemical Space with Machine Learning-Enabled Computational Calculations, *ACS Appl. Mater. Interfaces*, 2023, **15**(45), 52427–52435, DOI: **10.1021/acsami.3c10798**.

83 M. De Luca and A. Polimeni, Electronic properties of wurtzite-phase InP nanowires determined by optical and magneto-optical spectroscopy, *Appl. Phys. Rev.*, 2017, **4**(4), 041102, DOI: **10.1063/1.5006183**.

84 D. Stone, X. Li, T. Naor, J. Dai, S. Remennik and U. Banin, Size and Emission Control of Wurtzite InP Nanocrystals Synthesized from $Cu_{3-x}P$ by Cation Exchange, *Chem. Mater.*, 2023, **35**(24), 10594–10605, DOI: **10.1021/acs.chemmater.3c02226**.

85 D. O. Scanlon and A. Walsh, Bandgap engineering of ZnSnP2 for high-efficiency solar cells, *Appl. Phys. Lett.*, 2012, **100**(25), 251911, DOI: **10.1063/1.4730375**.

86 K.-J. Range and H.-J. Hübner, Die Kristallstruktur von MnIn2Te4, eine neue Ordnungsvariante für Defektzinkblendephasen/The Crystal Structure of MnIn2Te4, a New Type of Ordered Defect Zincblende Phases, *Z. Naturforsch. B*, 1975, **30**(3–4), 145–148, DOI: **10.1515/znb-1975-3-401**.

87 P. Dotzel, E. Franke, H. Schäfer and G. Schön, Zur Kenntnis von MgAl2Te4, MgGa2Te4 und MgIn2Te4/On the Ternary Tellurides MgAl2Te4, MgGa2Te4 and MgIn2Te4, *Z. Naturforsch. B*, 1975, **30**(3–4), 179–182, DOI: **10.1515/znb-1975-3-409**.

88 J.-X. Zhang, M.-Y. Ran, X.-T. Wu, H. Lin and Q.-L. Zhu, An overview of Mg-based IR nonlinear optical materials, *Inorg. Chem. Front.*, 2023, **10**(18), 5244–5257, DOI: **10.1039/D3QI01144E**.

89 P. Wang, Y. Chu, A. Tudi, C. Xie, Z. Yang, S. Pan and J. Li, The Combination of Structure Prediction and Experiment for the Exploration of Alkali-Earth Metal-Contained Chalcopyrite-Like IR Nonlinear Optical Material, *Adv. Sci.*, 2022, **9**(15), 2106120, DOI: **10.1002/advs.202106120**.

90 Y. Song, M. Luo, D. Zhao, G. Peng, C. Lin and N. Ye, Explorations of new UV nonlinear optical materials in the Na2CO3–CaCO3 system, *J. Mater. Chem. C*, 2017, **5**(34), 8758–8764, DOI: **10.1039/C7TC02789C**.

91 B. Zhang, G. Shi, Z. Yang, F. Zhang and S. Pan, Fluorooxoborates: Beryllium-Free Deep-Ultraviolet Nonlinear Optical Materials without Layered Growth, *Angew. Chem., Int. Ed.*, 2017, **56**(14), 3916–3919, DOI: **10.1002/anie.201700540**.

92 C. W. Glass, A. R. Oganov and N. Hansen, USPEX---Evolutionary crystal structure prediction, *Comput. Phys. Commun.*, 2006, **175**(11), 713–720, DOI: **10.1016/j.cpc.2006.07.020**, ISSN 0010-4655.

93 J. Noh, G. H. Gu, S. Kim and Y. Jung, Machine-enabled inverse design of inorganic solid materials: promises and challenges, *Chem. Sci.*, 2020, **11**(19), 4871–4881, DOI: **10.1039/D0SC00594K**, ISSN 2041-6520.

94 Z. Ren, S. Isaac Parker Tian, J. Noh, F. Oviedo, G. Xing, J. Li, Q. Liang, R. Zhu, A. G. Aberle, S. Sun, X. Wang, Y. Liu, Q. Li, S. Jayavelu, K. Hippalgaonkar, Y. Jung and T. Buonassisi, An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties, *Matter*, 2022, **5**(1), 314–335, DOI: **10.1016/j.matt.2021.11.032**, ISSN 2590-2393.

95 Q. Wu, L. Dong, L. Kang and Z. Lin, Prediction and Evaluation of Li2ZnS2 Crystals as Mid-Infrared Nonlinear Optical Material with High Thermal Conductivity, *Adv. Opt. Mater.*, 2025, 2402922, DOI: **10.1002/adom.202402922**, ISSN 2195-1071.

96 Q. Liu, R. An, C. Li, D. Chu, W. Zhao, S. Pan and Z. Yang, Accelerating Discovery of Infrared Nonlinear Optical Materials with High Lattice Thermal Conductivity: Combining Machine Learning and First-Principles Calculations, *Adv. Opt. Mater.*, 2025, 2403292, DOI: **10.1002/adom.202403292**, ISSN 2195-1071.

97 W. Zhang, H. Yu, H. Wu and P. Shiv Halasyamani, Phase-Matching in Nonlinear Optical Compounds: A Materials Perspective, *Chem. Mater.*, 2017, **29**(7), 2655–2668, DOI: **10.1021/acs.chemmater.7b00243**, ISSN 0897-4756.

98 Y. Lou and A. M. Ganose, Discovery of highly anisotropic dielectric crystals with equivariant graph neural networks, *Faraday Discuss.*, 2025, **256**, 255–274, DOI: **10.1039/D4FD00096J**, ISSN 1359-6640.

