Fast predictions of lattice energies by continuous isometry invariants of crystal structures*

Jakob Ropers[1], Marco M Mosca[1], Olga Anosova[1][0000−0003−4134−4398], Vitaliy Kurlin[1][0000−0001−5328−5351], and Andrew I Cooper[1]

University of Liverpool, Liverpool L69 3BX, UK vkurlin@liv.ac.uk
http://kurlin.org

Abstract. Crystal Structure Prediction (CSP) aims to discover solid crystalline materials by optimizing periodic arrangements of atoms, ions or molecules. CSP takes weeks of supercomputer time because of slow energy minimizations for millions of simulated crystals. The lattice energy is a key physical property, which hints at thermodynamic stability of a crystal but has no simple analytic expression. Past machine learning approaches to predict the lattice energy used slow crystal descriptors depending on manually chosen parameters. The new area of Periodic Geometry offers much faster isometry invariants that are also continuous under perturbations of atoms. Our experiments on simulated crystals confirm that a small distance between the new invariants guarantees a small difference of energies. We compare several kernel methods for invariant-based predictions of energy and achieve the mean absolute error of less than 5kJ/mole or 0.05eV/atom on a dataset of 5679 crystals.

Keywords: crystal · energy · isometry invariant · machine learning

1 Motivations, problem statement and overview of results

Solid crystalline materials (*crystals*) underpin key technological advances from solid-state batteries to therapeutic drugs. Crystals are still discovered by trial and error in a lab, because their properties are not yet expressed in terms of crystal geometries. This paper makes an important step towards understanding the structure-property relations, for example how an energy of a crystal depends on its geometric structure. The proposed methods belong to the recently established area of Periodic Geometry, which studies geometric descriptors (*continuous isometry invariants*) and metrics on a space of all periodic crystals.

The most important property of a crystal is the energy of its crystal structure, which is usually called the *lattice energy* or *potential energy surface* or *energy landscape* [29]. This lattice energy hints at thermodynamic stability of a crystal, whether such a crystal can be accessible for synthesis in a lab and can remain stable under application conditions. Since the lattice energy has no

* Supported by £3.5M EPSRC grant 'Application-driven Topological Data Analysis'.

closed analytic expression, calculations are always approximate, from the *force field* (FF) level [19] to the more exact density functional theory (DFT) [12].

Our experiments use the lattice energy obtained by force fields for the CSP data of 5679 nanoporous T2 crystals in Fig. 1 predicted by our colleagues [23].

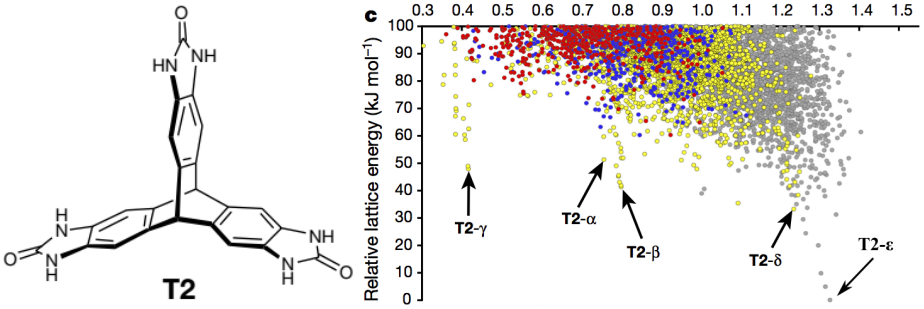

Fig. 1. Left: T2 molecule. **Right**: energy-vs-density plot of 5679 predicted crystals, five polymorphs were synthesized, most recent T2-ε crystal is added to [23, Fig. 2d].

Traditionally a periodic crystal is stored in a Crystallographic Information File (CIF). This file specifies a linear basis v_1, v_2, v_3 of $\mathbb{R}^3$, which spans the *unit cell* $U = \{\sum_{i=1}^{3} c_i v_i \mid 0 \le c_i < 1\}$, generates the *lattice* $\Lambda = \{\sum_{i=1}^{3} c_i v_i \mid c_i \in \mathbb{Z}\}$. Then a crystal can be obtained as the infinite union of lattice translates $M + \Lambda = \{p + v : p \in M, v \in \Lambda\}$ from a finite set (*motif*) of points $M \subset U$ in the cell U. The representation $M + \Lambda$ is simple but is highly ambiguous in the sense that infinitely many pairs (cell, motif) generate equivalent crystals, see [3, Fig. 2].

The main novelty of our approach to energy predictions is using the fast computable and easily interpretable invariants of crystals. The concept of an invariant has a rigorous definition after we fix an equivalence relation on objects in question. Since crystal structures are determined in a rigid form, the most natural equivalence is rigid motion or *isometry*, which is any map that preserves interpoint distances, for example a composition of translations and rotations in Euclidean space $\mathbb{R}^3$. Any orientation-preserving isometry can be realized as a *rigid motion*, which is a continuous family f_t, $t \in [0, 1]$, of isometries starting from the identity map $f_0 = \mathrm{id}$. Since any general isometry is a composition of a single reflection and a rigid motion, we consider isometry as our main *equivalence relation* on crystals. Later we can also take into account a sign of orientation.

An *isometry invariant* I is a crystal property or a function, say from crystals to numbers, preserved by isometry. So if crystals S, Q are isometric then $I(S) = I(Q)$. The classical example invariants of a crystal S are the symmetry group (the group of isometries that map S to itself) and the volume of a minimal (*primitive*) unit cell. Example non-invariants are unit cell parameters (edge-lengths and angles) and fractional coordinates of atoms in a cell basis.

Many widely used isometry invariants including symmetry groups break down (are *discontinuous*) under perturbations of atoms, which always exist in real crystals at finite temperature. Perturbations are also important for distinguishing simulated crystals obtained via Crystal Structure Prediction (CSP). Indeed, CSP iteratively minimizes the lattice energy and inevitably stops at some approximation to a local minimum [20]. Hence, after many random initializations, we likely get many near duplicate structures around the same local minumum.

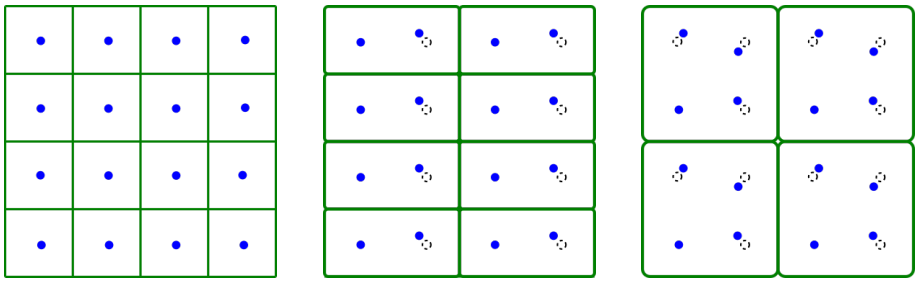

Fig. 2. Most past invariants are discontinuous under perturbations above, for example symmetry groups and sizes of primitive or reduced cells. Recent isometry invariants [17, 31, 32] continuously quantify similarities between perturbed periodic structures.

Since any perturbation of points keeping their periodicity (but not necessarily an original unit cell) produces a new close structure, all periodic structures form a continuous space. Then any CSP dataset can be viewed as a discrete sample from the underlying continuous space of periodic structures. The lattice energy is a function on this crystal space whose geometry needs to be understood. The problem below is a key step towards describing structure-property relations.

Properties-from-invariants problem. Find suitable isometry invariants that justifiably predict desired properties of crystals such as the lattice energy. ■

The proposed invariants to tackle the above problem are average minimum distances (AMD) [31]. AMD is an infinite sequence of isometry invariants whose values change by at most 2ε if given points are perturbed in their ε-neighborhoods. A thousand of AMD invariants can be computed in milliseconds on a modest desktop for crystals with hundreds of atoms in a unit cell [31, appendix D].

The above continuity of AMD guarantees that perturbed crystals have close AMD values. Then such a theoretically continuous invariant can be tested for checking continuity of energy under crystal perturbations. The first contribution is an experimental detection of constants λ and δ such that, for any smaller distance $d < \delta$ between AMD vectors, the corresponding crystals have a lattice energy difference within λd, usually within 2kJ/mole. Past invariants have no such a constant to quantify continuity of energy in this way. For example, close values of density, RMSD [8], PXRD [25] don't guarantee close values of energy.

The second contribution is the demonstration that several kernel methods can achieve a mean absolute error of less than 5kJ/mole by using only isometry invariants without any chemical data. The key achievement is the time of less than 10 min for training by using a modest desktop on a dataset of 5679 structures, while energy predictions take milliseconds per crystal on average.

Section 2 reviews closely related past work using crystal descriptors for machine learning of the lattice energy. Section 3 reminds the recently introduced isometry invariants of periodic point sets and their properties. Section 4 quantifies continuity of energy in terms of AMD invariants. Section 5 describes how the energy of a crystal can be predicted from its AMD invariants by using several kernel methods. Section 6 discusses limitations and potential developments.

2 Review of related machine learning approaches

This section reviews the closest related work about energy predictions for infinite periodic crystals. The same problem is simpler for a single molecule [27].

Energy predictions use various representations of crystals. We review only geometric descriptors that are closest to isometry invariants in section 3.

The partial radial distribution function (PRDF) is based on the density of atoms of type β in a shell of radius r and width dr centered around an atom of type α [26]. Since atom types are essentially used, the PRDF can be best for comparing crystals that are composed of the same atom types. Due to averaging across all atoms of a type α within a unit cell, the PRDF is independent of a cell choice. A similar distance-based fingerprint was introduced earlier by Valle and Oganov [28]. Since only pairwise distances are used, these descriptors are isometry invariants and likely continuous under perturbations shown in Fig. 2.

Completeness or uniqueness of a crystal with a given PRDF is unclear yet, but can be theoretically possible for a large enough radius r. Practical computations require choices or the distance thresholds r and dr, which can affect the PRDF. Schutt et al. confirm in [26, Table I] that the PRDF outperforms non-invariant features such as the Bravais matrix of cell parameters. The mean absolute error (MAE) of energy predictions based on PRDF is 0.68eV/atom or 65.6kJ/mole.

Another way to build geometric attributes of a crystal structure is to use Wigner-Seitz cells (also called Dirichlet or Voronoi domains) of atoms. Ward et al. [30] used 271 cell-based geometric and chemical attributes to achieve the MAE of 0.09eV/atom or 8.7kJ/mole for predicting the formation enthalpy.

An extensible neural network potential [27, Fig. 4] has further improved the mean absolute error (MAE) to 1.8kcal/mole=7.56kJ/mole. The most advanced approach by Egorova et al. [10] predicts the difference between the accurate DFT energy and its force field approximation a with MAE less than 2kJ/mole by using GGA DFT (PBE) calculations and symmetry function descriptors [5].

3 Key definitions and recent results of Periodic Geometry

This section reviews more recent work in the new area of Periodic Geometry [2], which studies the metric geometry on the space of all periodic structures. Nuclei of atoms are better defined physical objects than chemical bonds, which depend on many thresholds for distances and angles. Hence the most fundamental model of a crystal is a periodic set of zero-sized points representing all atomic centers.

Though chemical elements and other physical properties can be easily added to invariants as labels of points, the experiments in [9, 31] and sections 4, 5 show that the new invariants can be enough to infer all chemistry from geometry.

The symbol $\mathbb{R}^n$ denotes Euclidean space with *Euclidean* distance $|p-q|$ between points $p, q \in \mathbb{R}^n$. Motivated by a traditional representation of a crystal by a Crystallographic Information File, a periodic point set S is given by a pair (cell U, motif M). Here U is a *unit cell* (parallelepiped) spanned by a linear basis $v_1, \dots, v_n$ of $\mathbb{R}^n$, which generates the *lattice* $\Lambda = \{\sum_{i=1}^{n} c_i v_i : c_i \in \mathbb{Z}\}$. A *periodic point set* $S = M + \Lambda$ is obtained by shifting a finite *motif* $M \subset U$ of points along all vectors $v \in \Lambda$. Fig. 3 illustrates the problem of transforming ambiguous input into invariants that can distinguish periodic sets up to isometry.

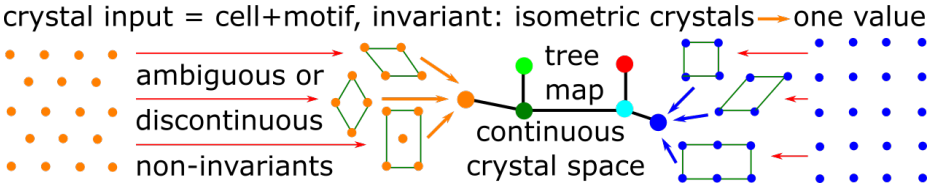

Fig. 3. Any periodic sets, for example the hexagonal and square lattices, can be represented by infinitely many pairs (cell, motif). This ambiguity can be resolved only by a complete isometry invariant that should continuously parameterize the crystal space.

Arguably the simplest isometry invariant of a crystal is its density ρ. Without distinguishing atoms in a periodic point set S, the density $\rho(S)$ is the number m of points in a unit cell U, divided by the cell volume $\mathrm{Vol}[U]$. The density ρ distinguishes hexagonal and square lattices in Fig. 3 but is insensitive to perturbations shown in Fig. 2. Though many real crystals are dense and can not be well-separated by density, energy landscapes are still visualized as energy-vs-density plots in Fig. 1. The single-value density ρ has been recently extended to the sequence of density functions $\psi_k(t)$ [9]. For any integer $k \geq 1$, the *density function* $\psi_k(t)$ measures the volume of the regions within a unit cell U covered by k balls with radius $t \geq 0$ and centers at all points $p \in M$, divided by $\mathrm{Vol}[U]$.

Though these isometry invariants have helped to identify a missing crystal in the Cambridge Structural Database, their running time cubically depends on k, which is a bit slow for big datasets. The following invariants are much faster.

Let a periodic point set $S \subset \mathbb{R}^n$ have points $p_1, \dots, p_m$ in a unit cell. For any $k \geq 1$ and $i = 1, \dots, m$, the i-th row of the $m \times k$ matrix $D(S;k)$ consists of the ordered distances $d_{i1} \leq \dots \leq d_{ik}$ measured from the point p_i to its first k nearest neighbors within the infinite set S, see Fig. 4. The *Average Minimum Distance* $\mathrm{AMD}_k(S) = \dfrac{1}{m} \sum\limits_{i=1}^{m} d_{ik}$ is the average of the k-th column in $D(S;k)$.

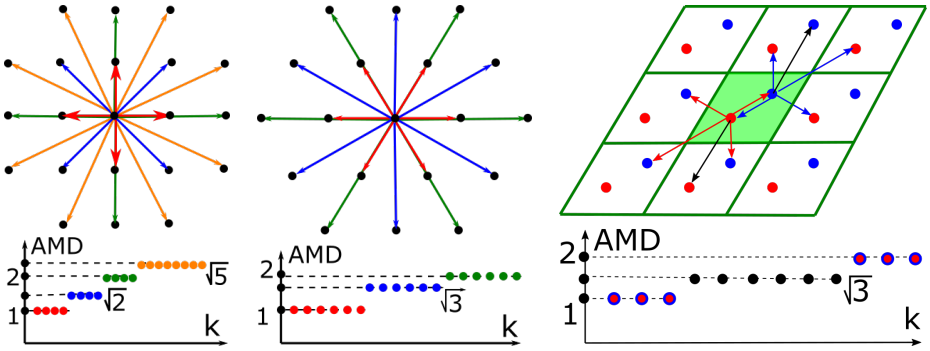

Fig. 4. [31, Fig. 4] **Left**: in the square lattice, distances from the origin to its first few neighbors are shown in the graph of AMD_k values, e.g. the shortest axis-aligned distances are $\mathrm{AMD}_1 = \dots = \mathrm{AMD}_4 = 1$, the longer diagonal distances are $\mathrm{AMD}_5 = \dots = \mathrm{AMD}_8 = \sqrt{2}$. **Middle**: in the hexagonal lattice, the shortest distances are $\mathrm{AMD}_1 = \dots = \mathrm{AMD}_6 = 1$. **Right**: AMD for a honeycomb periodic set (graphene).

[31, Theorem 4] proves that AMD is an isometry invariant independent of a unit cell. The AMD invariants are similar to radial distribution functions [26] and related density-based invariants [28]. The AMD definition has no manually chosen thresholds such as cut-off radii or tolerances. The length k of the vector $\mathrm{AMD}^{(k)} = (\mathrm{AMD}_1, \dots, \mathrm{AMD}_k)$ is not a parameter in the sense that increasing k only adds new values without changing previous ones. Hence k can be considered as an order of approximation, similarly to an initial length of a DNA code.

We have no examples of non-isometric sets that have identical infinite AMD sequences. Hence AMD can be complete at least for periodic sets in general position so that if two sets S, Q have $\mathrm{AMD}(S) = \mathrm{AMD}(T)$, then S, Q are isometric. More recently, the isometry classification of all periodic point sets was reduced to an *isoset* [3], which is a collection of atomic environments considered modulo rotations and up to a *stable* radius α. This stable radius is defined for a given crystal and any two crystals can be compared by isosets of their maximum radius so that two sets S, Q are isometric if and only if their isosets are equivalent.

This paper uses AMD invariants due to their easy interpretability and fast running time. $\mathrm{AMD}_k(S)$ asymptotically approaches $c(S)\sqrt[n]{k}$, where $c(S)$ is related to the density of a periodic point set $S \subset \mathbb{R}^n$, see [31, Theorem 13]. A near linear computational time [31, Theorem 14] of AMD_k in both m, k translates into milliseconds on a modest laptop, which allowed us to visualize all 229K organic molecular crystals from the Cambridge Structural Database in a few hours.

4 Continuity of the energy in terms of AMD invariants

To express continuity of AMD and other invariants under perturbations, we use the maximum displacement of atoms formalized by the *bottleneck distance* d_B as follows. For any bijection $g : S \to Q$ between periodic point sets, the maximum displacement is $d_g(S,Q) = \sup_{p \in S} |g(p) - p|$. After minimizing over all bijections $g : S \to Q$, we get the *bottleneck distance* $d_B(S,Q) = \inf_{g:S \to Q} d_g(S,Q)$.

The structure-property hypothesis says that all properties of a crystal should be determined by its geometric structure. Understanding how any property can be explicitly computed from a crystal structure would replace trial-and-error methods by a guided discovery to find crystals with desired properties.

Most current attempts are based on black-box machine learning of properties from crystal descriptors, not all of which are invariants up to isometry. All machine learning tools rely on the usually implicit assumption that small perturbations in input data lead to relatively small perturbations in outputs.

Continuity of a structure-property relation can be mathematically expressed as Lipschitz continuity [21, section 9.4]: $|E(S) - E(Q)| \le \lambda d(S,Q)$, where λ is a constant, E is a crystal property such as the lattice energy, $d(S,Q)$ is a distance satisfying all metric axioms on crystals S, Q or their invariants. The above inequality should hold for all crystals S, Q with small distances $d(S,Q) < \delta$, where a threshold δ may depend on a property E or a metric d, not on S, Q.

The continuity above sounds plausible and seems necessary for the structure-property hypothesis. Indeed, if even small perturbations of a geometric structure drastically change crystal properties, then any inevitably noisy structure determination would not suffice to guarantee desired properties of a crystal.

Fig. 5,6,7 show that the past methods of characterizing crystal similarity are insufficient to guarantee the above continuity of the lattice energy. These results were obtained on the T2 dataset of 5679 simulated crystals reported in [23]. Each square dot represents a pair of crystals with differences in past descriptors on the horizontal axis and differences in energies on the vertical axis.

Fig. 5 shows dozens of crystal pairs with very close densities and rather different lattice energies, which means that the energy discontinuously varies relative to the density. This failure of a single-value descriptors might not be surprising not only for crystals, which are often very dense materials, but also for other real-life scenarios. For example, many people have the same height and very different weights. However, the density is still used to represent a crystal structure in CSP landscapes such as Fig. 1. Indeed, the density is an isometry invariant, which is continuous (actually, constant) under perturbations, see Fig. 2.

Fig. 6 illustrates a similar conclusion for the traditional packing similarity measured by the COMPACK algorithm [8] as the Root Mean Square Deviation (RMSD) of atomic positions matched between up to 15 (by default) molecules in

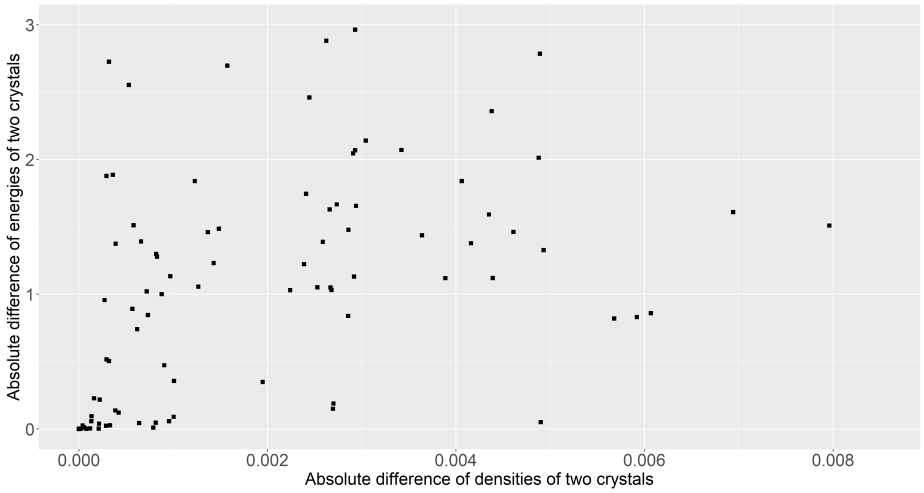

Fig. 5. 5679 crystals in Fig. 1 have the density range $[0.3, 1.4]$. Many crystals have differences in densities within 0.003g/cm^3 and differences in energies up to 3kJ/mole.

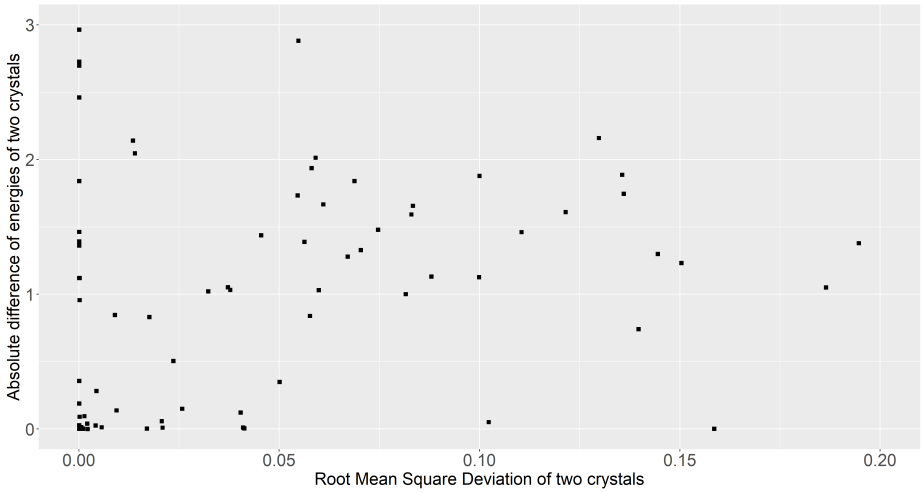

Fig. 6. Crystal pairs with RMSD $< 0.1\mathring{A}$ have differences in energies up to 3kJ/mole.

two crystals. This similarity relies on two extra thresholds for atomic distances and angles whose values affect the RMSD. For example, when only one of 15 molecules is matched, the RMSD is exactly 0, because all 5679 crystals are based on the same T2 molecule in Fig. 1. Nonetheless, this packing similarity can visually confirm that nearly identical crystals nicely overlap each other.

The powder X-ray diffraction (PXRD) similarity has the range [0,1] with values close to 1 indicating closeness of diffraction patterns. Fig. 7 has $1-$PXRD

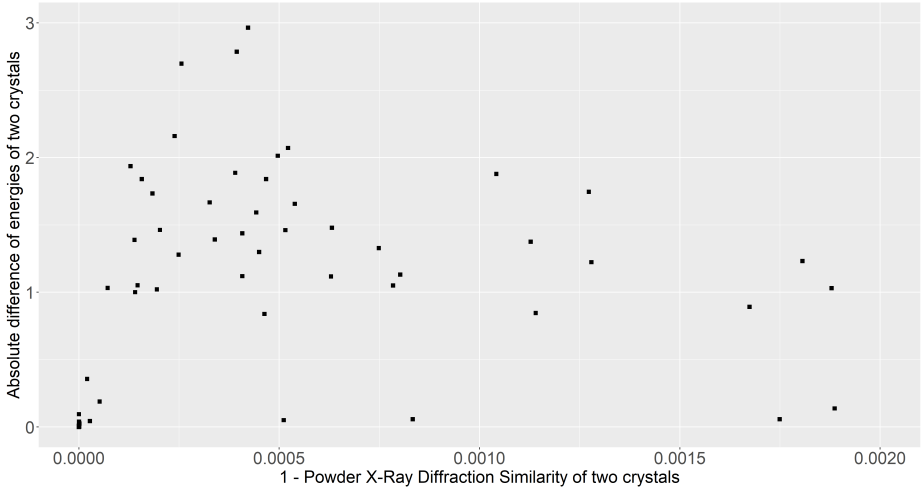

Fig. 7. Crystal pairs with PXRD similarity > 0.9995 have big differences in energies.

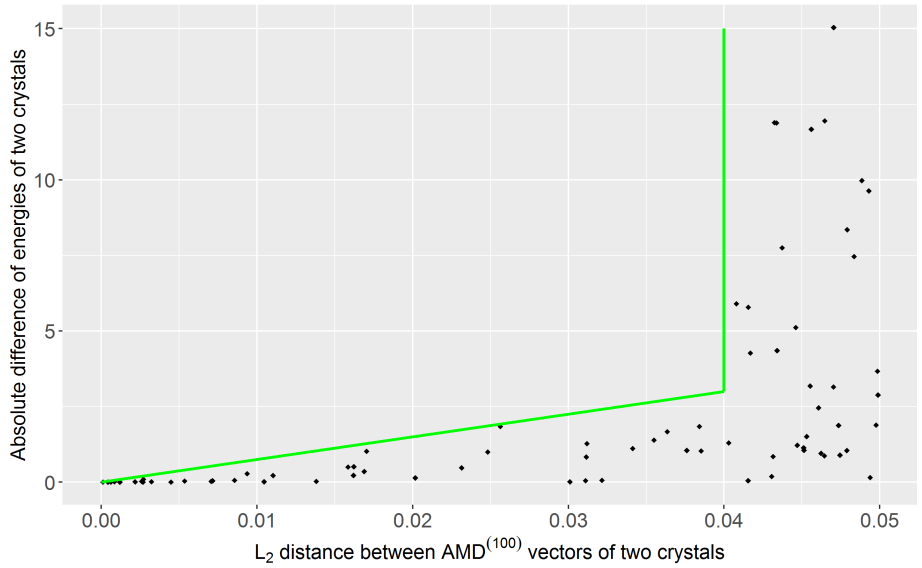

Fig. 8. The green line $|\Delta E| = 75L_2$ over $L_2 \in [0, 0.04]$ shows that if crystals have a distance $L_2 < 0.04\mathring{A}$ between $\text{AMD}^{(100)}$ vectors, their energies differ by at most $75L_2$.

on the horizontal axis and similarly to the above two plots shows many pairs of nearly identical crystals (with PXRD above 0.9995) with rather different energies. Despite Fig. 5,6,7 illustrating the discontinuity of the lattice energy with respect to traditional similarity measures of crystals, we should not despair.

The new AMD invariants detect tiny differences in crystal structures and are continuous under perturbations in the bottleneck distance [31, Theorem 9]: $|\mathrm{AMD}_k(S) - \mathrm{AMD}_k(Q)| \le 2d_B(S,Q)$ if the bottleneck distance d_B is less than half of the minimum distance between points in any of periodic sets $S, Q \subset \mathbb{R}^n$.

Even more importantly, Fig. 8, 9,10 show that the lattice energy continuously changes with respect to AMD invariants on the same T2 dataset. Each rhombic dot in Fig. 8,9,10 represents one pairwise comparison between $\mathrm{AMD}^{(100)}$ vectors of length $k = 100$ for two T2 crystals. The distances between vectors $p = (p_1, \dots, p_k)$ and $q = (q_1, \dots, q_k)$ on the horizontal axis are computed by the Euclidean metric $L_2(p,q) = \sqrt{\sum\limits_{i=1}^{k} |p_i - q_i|^2}$, the Chebyshev metric $L_\infty(p,q) = \max\limits_{i=1,\dots,k} |p_i - q_i|$ and the Manhattan metric $L_1(p,q) = \sum\limits_{i=1}^{k} |p_i - q_i|$.

Despite the T2 dataset being thoroughly filtered out to remove near duplicates, Fig. 8, 9 include several pairs whose AMD invariants are very close, though not identical. In all these cases the corresponding crystals also have very close energies, which can be quantified via Lipschitz continuity as follows.

In Fig. 8 the Lipschitz continuity for the energy $|\Delta E| = |E(S) - E(Q)| \le \lambda_2 L_2(\mathrm{AMD}^{(100)}(S), \mathrm{AMD}^{(100)}(Q))$ holds for $\lambda_2 = 200$ and all pairs of crystals S, Q whose $\mathrm{AMD}^{(100)}$ vectors have a Euclidean distance $L_2 < \delta_2 = 0.04\mathring{A}$.

Visually, all these pairs are below the green line $\Delta E = 200L_2$ up to the distance threshold $\delta_2 = 0.04\mathring{A}$. If distances between crystals become too large, a single-value metric cannot guarantee close values of energy. Using the geographic analogy, the further we travel from any fixed location on planet Earth, the more variation in physical properties such as the altitude we should expect.

Fig. 9 illustrates continuity of the lattice energy with respect to the metric $L_\infty(p,q) = \max\limits_{i=1,\dots,k} |p_i - q_i|$ between $\mathrm{AMD}^{(100)}$ vectors. All pairs of crystals with distances $L_\infty < \delta_\infty = 0.009\mathring{A}$ have differences in energies less than $\lambda_\infty L_\infty$ with $\lambda_\infty = 200$, so all corresponding dots are below the green line $|\Delta E| = 200L_\infty$.

Fig. 10 shows that the lattice energy continuously behaves for the metric $L_1(p,q) = \sum\limits_{i=1}^{k} |p_i - q_i|$ between $\mathrm{AMD}^{(100)}$ vectors. All pairs of crystals with distances $L_1 < \delta_1 = 0.32\mathring{A}$ have energy differences less than $\lambda_1 L_1$ with $\lambda_1 = 10$, so all corresponding dots are below the green line $|\Delta E| = 10L_1$.

The thresholds $\delta_1 = 0.32$ and $\delta_2 = 0.04$ are larger than $\delta_\infty = 0.009\mathring{A}$, because the metrics L_1, L_2 sum up all deviations between corresponding coordinates of $\mathrm{AMD}^{(100)}$ vectors, while the metric L_∞ measures only the maximum deviation.

If we tried to fit Lipschitz continuity for the past descriptors (density, RMSD, PXRD) in Fig. 5,6,7 similarly to AMD invariants above, corresponding green lines would be almost vertical with huge slopes or gradients (Lipschitz constants).

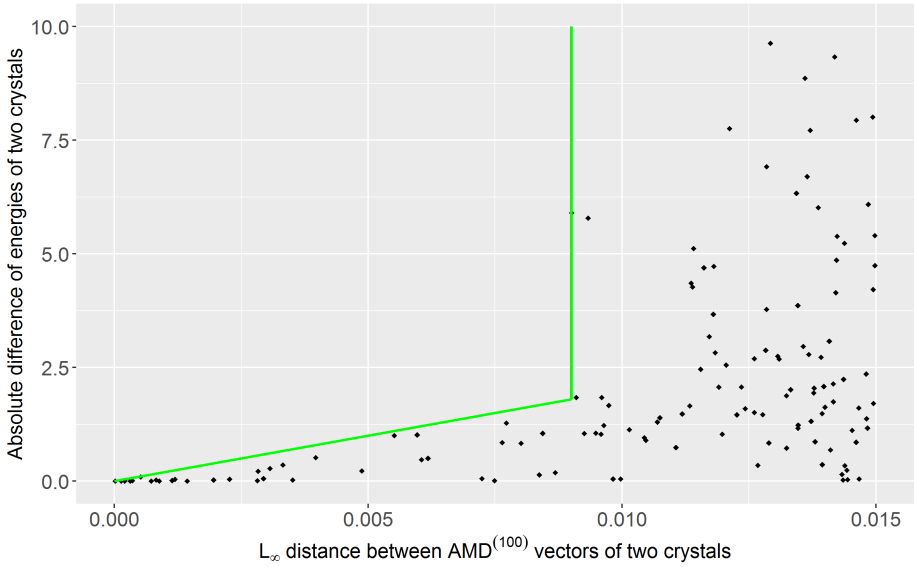

Fig. 9. The green line $|\Delta E| = 200L_\infty$ over $[0, 0.009]$ shows that if crystals have a distance $L_\infty < 0.009\mathring{A}$ between $\text{AMD}^{(100)}$, their energies differ by at most $200L_\infty$.

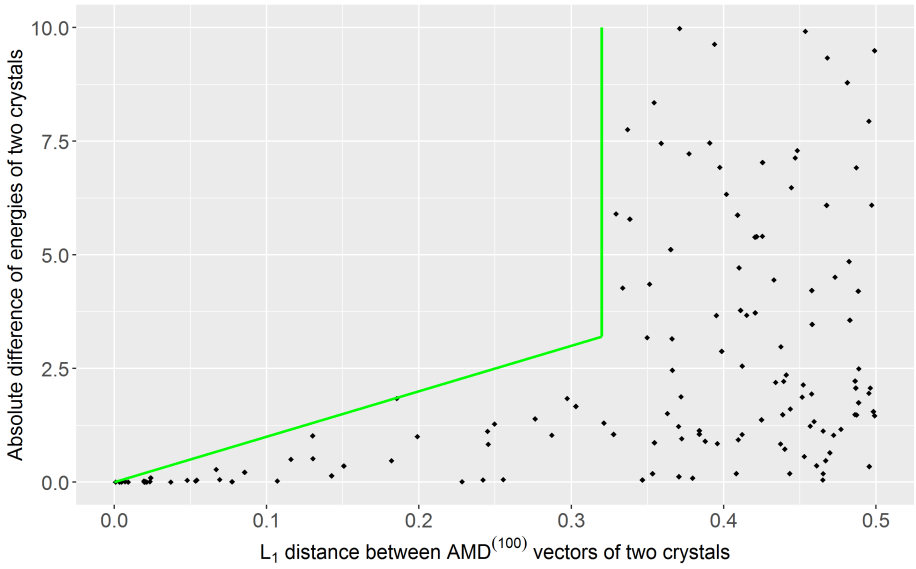

Fig. 10. The green line $|\Delta E| = 10L_1$ over $L_1 \in [0, 0.32]$ shows that if crystals have a distance $L_1 < 0.32\mathring{A}$ between $\text{AMD}^{(100)}$ vectors, their energies differ by at most $10L_1$.

5 Fast predictions of the energy by AMD invariants

This section describes the second important contribution by showing that continuity of AMD from section 4 leads to state-of-the-art energy predictions.

The energy prediction problem is to infer the lattice energy from a crystal structure, for example by using a dataset of ground truth energies for training.

The descriptors in Fig. 5,6,7 cannot be justifiably used to resolve the above problem because of their discontinuity. Indeed, if we input a slightly different (say, experimental) crystal, we expect a close value of energy in the output.

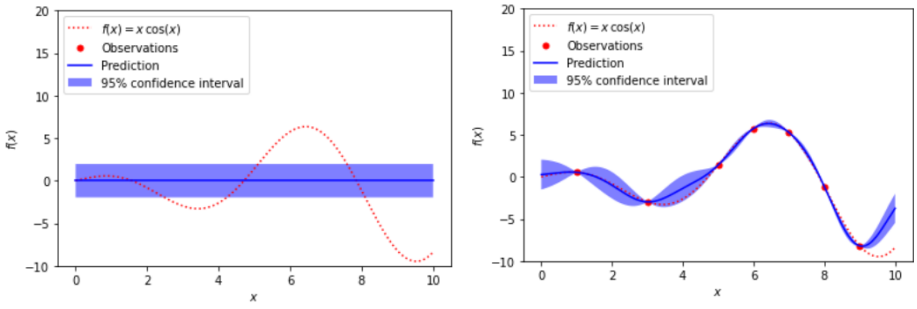

Fig. 11. Gaussian Process tries to predict values of $f(x) = x\cos x$ by training on observed data points. **Left**: an initial prediction is 0 for any x. **Right**: predictions substantially improve after training on six data points under the natural assumption that the underlying function is *continuous*, so continuity is important for learning.

First we describe the Gaussian Process Regression [13] as implemented in SciKit Learn [22], see Fig. 11, which achieved the best results on the T2 dataset of 5679 crystals. Initially each T2 crystal is converted into a periodic point set S by placing a zero-sized point at every atomic center. Then each S is represented by its $\mathrm{AMD}^{(k)}(S)$ vector of a fixed length in the range $k = 50, 100, \dots, 500$. The base distance d between $\mathrm{AMD}^{(k)}$ vectors was chosen as L_∞ due to the smallest Lipschitz constant $\lambda = 2$ in the continuity property $|\mathrm{AMD}_k(S) - \mathrm{AMD}_k(Q)| \le \lambda d_B(S,Q)$. For the metrics L_1, L_2, the Lipschitz constants would be $2k, 2\sqrt{k}$.

For any pair of crystals S, Q, we consider the Rational Quadratic Kernel $K(S,Q) = \left(1 + \dfrac{d^2(S,Q)}{2\alpha l^2}\right)^{-\alpha}$, where α, l are scale parameters optimized by training. For a single prediction run, the whole T2 dataset was randomly split into 80% training subset and remaining 20% test subset of $m = 1136$ crystals.

Table 1 shows three types of errors, each averaged over 10 runs above: $\mathrm{RMSE} = \sqrt{\dfrac{1}{m}\sum\limits_{i=1}^{m} |E_{true}(S_i) - E_{pred}(S_i)|^2}$ is the root mean square error in the

lattice energy averaged over m crystals $S_1, \dots, S_m$ from the test subset, then $\text{MAE} = \frac{1}{m} \max\limits_{i=1,\dots,m} |E_{true}(S_i) - E_{pred}(S_i)|$ is the mean absolute error and the mean absolute percentage error $\text{MAPE} = \frac{1}{m} \max\limits_{i=1,\dots,m} \frac{|E_{true}(S_i) - E_{pred}(S_i)|}{E_{true}(S_i)}$. Each value has the empirical standard deviation $\pm$std computed over 10 runs.

Table 1. The Gaussian Process with the Rational Quadratic Kernel predicts the energy reported in [23] with the mean absolute error (MAE) of less than 5kJ/mole on $m = 1136$ crystals by training on the isometry invariants $\text{AMD}^{(k)}$ of 4543 crystals for various k.

k	RMSE $\pm$ std	MAE $\pm$ std	MAPE $\pm$ std	training time, sec	full test time, ms
50	6.503 ± 0.123	4.900 ± 0.86	3.509 ± 0.059	627 ± 85	15961 ± 183
100	6.344 ± 0.152	4.801 ± 0.103	3.439 ± 0.070	349 ± 47	7979 ± 564
150	6.607 ± 0.119	4.977 ± 0.077	3.559 ± 0.053	400 ± 23	12789 ± 203
200	6.617 ± 0.147	4.966 ± 0.114	3.554 ± 0.079	506 ± 40	15943 ± 46
250	6.517 ± 0.109	4.914 ± 0.082	3.514 ± 0.055	574 ± 91	16464 ± 193
300	6.632 ± 0.139	5.003 ± 0.092	3.577 ± 0.062	545 ± 15	16431 ± 52
350	6.615 ± 0.077	4.990 ± 0.077	3.581 ± 0.053	500 ± 22	12395 ± 44
400	6.611 ± 0.149	4.984 ± 0.080	3.569 ± 0.053	585 ± 25	17906 ± 201
450	6.559 ± 0.179	4.954 ± 0.127	3.545 ± 0.085	512 ± 21	12927 ± 67
500	6.622 ± 0.116	5.004 ± 0.092	3.581 ± 0.068	598 ± 24	18429 ± 219

Table 1 shows that the errors RMSE, MAE, MAPE are consistent across different values of k. The key advantage over past methods is the speed: less than 10 min for training for 4543 vectors $\text{AMD}^{(k)}$ on Intel Xeon CPU at 2.3 GHz. The last column shows the full test time on $m = 1136$ crystals, so the average time per crystal is more than 1000 times faster. The smallest mean absolute error MAE $\approx$ 4.8kJ/mole corresponds to about 7.4 milliseconds (ms) per crystal.

The computation of $\text{AMD}^{(k)}$ asymptotically has a near linear time in k and the number of atoms in a unit cell by [31, Theorem 14], which needs only 27ms on average per T2 crystal for $k = 1000$ on a similar desktop. This ultra-fast speed allowed us to visualize for the first time all 229K molecular organic crystals from the Cambridge Structural Database in less than 9 hours, see [31, appendix D].

We have tried other types of kernels: the matern and linear kernels gave slightly larger errors, the squared exponential was worse for some k. We also considered another version of the T2 dataset without hydrogens (32 atoms per molecule instead of 46), which gave a bit bigger error for all kernels above.

Instead of AMD invariants, we trained the Gaussian Process Regression on the density functions $\psi_k(t)$ [9], which are continuous isometry invariants extending the single-value density for a variable radius $t \geq 0$. The average errors of AMD-based predictions were smaller than for the density functions ψ_k, which are also slower to compute than AMD, asymptotically in a cubic time in k.

Finally, the Random Forest [18] and Dense Neural Network [11] trained on AMD and density functions performed slight worse than the Gaussian Process, though the training and text times were much faster (seconds instead of minutes). The experiments above are reported in the dissertation of the first author [24].

6 Conclusions and a discussion of future developments

This paper has demonstrated that the recently developed continuous isometry invariants can provide insights undetected by traditional similarity measures.

In section 4 Fig. 5,6,7 show that many crystals can have almost identical density, RMSD, PXRD patterns but rather different lattice energies. On the same T2 dataset [23] Fig. 8,9,10 show that the lattice energy satisfies the Lipschitz continuity $|E(S) - E(Q)| \leq \lambda d(S,Q)$ for a fixed constant λ and all crystals S, Q whose AMD invariants are close with respect to the metrics L_1, L_2, L_∞.

In section 5 the standard kernel methods trained only on 100 isometry invariants $\mathrm{AMD}^{(100)}$ achieved the state-of-the-art mean absolute error of less than 5kJ/mole in energy. The key achievement is the speed of training (about 10 min for 4543 crystals on a modest desktop) and testing, which run in milliseconds per crystal. The code of experiments in section 5 is available on GitHub [24].

It should not be surprising that the lattice energy can be efficiently predicted from distance-based invariants without any chemical information. Indeed, if one atom is replaced by a different chemical element, then inter-atomic distances to neighbors inevitably change, even if slightly. These differences in distances can be detected, also after averaging over motif points. So AMD should pick up differences in crystals after swapping different atoms. The recent papers [1, 16, 15, 7, 14, 6, 4] defined complete invariants in partial cases and Pointwise Distance Distributions [32] detecting unexpected duplicates in the CSD [31, section 7].

References

1. Anosova, O., Kurlin, V.: Density functions of periodic sequences. arxiv:2205.02226
2. Anosova, O., Kurlin, V.: Introduction to periodic geometry and topology. arXiv:2103.02749 (2021)
3. Anosova, O., Kurlin, V.: An isometry classification of periodic point sets. In: Proceedings of Discrete Geometry and Mathematical Morphology (2021)
4. Anosova, O., Kurlin, V.: Algorithms for continuous metrics on periodic crystals. arXiv:2205.15298 (2022)
5. Behler, J.: Atom-centered symmetry functions for constructing high-dimensional neural network potentials. The Journal of chemical physics **134**(7), 074106 (2011)

6. Bright, M., Cooper, A.I., Kurlin, V.: Welcome to a continuous world of 3-dimensional lattices. arxiv:2109.11538
7. Bright, M., Cooper, A.I., Kurlin, V.: Geographic-style maps for 2-dimensional lattices. arXiv:2109.10885 (2021)
8. Chisholm, J., Motherwell, S.: Compack: a program for identifying crystal structure similarity using distances. J. Applied Crystallography **38**(1), 228–231 (2005)
9. Edelsbrunner, H., Heiss, T., Kurlin, V., Smith, P., Wintraecken, M.: The density fingerprint of a periodic point set. In: Proceedings of SoCG (2021)
10. Egorova, O., et al.: Multifidelity statistical machine learning for molecular crystal structure prediction. J Phys. Chemistry A **124**, 8065–8078 (2020)
11. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning, vol. 1. MIT Press (2016)
12. Gross, E., Dreizler, R.: Density functional theory, vol. 337 (2013)
13. KI Williams, C.: Gaussian processes for machine learning. Taylor & Francis (2006)
14. Kurlin, V.: A complete isometry classification of 3-dimensional lattices. arxiv:2201.10543
15. Kurlin, V.: Mathematics of 2-dimensional lattices. arxiv:2201.05150
16. Kurlin, V.: A computable and continuous metric on isometry classes of high-dimensional periodic sequences. arxiv:2205.04388 (2022)
17. Mosca, M., Kurlin, V.: Voronoi-based similarity distances between arbitrary crystal lattices. Crystal Research and Technology **55**(5), 1900197 (2020)
18. Myles, A., et al.: An introduction to decision tree modeling. J Chemometrics **18**(6), 275–285 (2004)
19. Niketic, S.R., Rasmussen, K.: The consistent force field: a documentation, vol. 3. Springer Science & Business Media (2012)
20. Oganov, A.: Modern methods of crystal structure prediction. Wiley & Sons (2011)
21. O'Searcoid, M.: Metric spaces. Springer Science & Business Media (2006)
22. Pedregosa, F., et al.: Scikit-learn: Machine learning in python. J Machine Learning Research **12**, 2825–2830 (2011)
23. Pulido, A., et al.: Functional materials discovery using energy–structure maps. Nature **543**, 657–664 (2017)
24. Ropers, J.: Applying machine learning to geometric invariants of crystals (2021), `https://github.com/JRopes/CrystalEnergyPrediction`
25. Sacchi, P., Lusi, M., Cruz-Cabeza, A.J., Nauha, E., Bernstein, J.: Same or different–that is the question: identification of crystal forms from crystal structure data. CrystEngComm **22**(43), 7170–7185 (2020)
26. Schütt, K., Glawe, H., Brockherde, F., Sanna, A., Müller, K.R., Gross, E.: How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. Physical Review B **89**(20), 205118 (2014)
27. Smith, J., Isayev, O., Roitberg, A.: An extensible neural network potential with dft accuracy at force field computational cost. Chem. Science **8**, 3192–3203 (2017)
28. Valle, M., Oganov, A.R.: Crystal fingerprint space–a novel paradigm for studying crystal-structure sets. Acta Crystallographica A **66**(5), 507–517 (2010)
29. Wales, D.: Exploring energy landscapes. Ann. Rev. Phys. Chem. **69**, 401–425 (2018)
30. Ward, L., Liu, R., Krishna, A., Hegde, V., Agrawal, A., Choudhary, A., Wolverton, C.: Including crystal structure attributes in machine learning models of formation energies via voronoi tessellations. Physical Review B **96**(2), 024104 (2017)
31. Widdowson, D., et al.: Average minimum distances of periodic point sets. MATCH Comm. Math. Comput. Chemistry **87**, 529–559 (2022)
32. Widdowson, D., Kurlin, V.: Pointwise distance distributions of periodic sets. arxiv:2108.04798