

# The Density Fingerprint of a Periodic Point Set

Herbert Edelsbrunner    Teresa Heiss    Vitaliy Kurlin    Philip Smith  
Mathijs Wintraecken

## Abstract

Modeling a crystal as a periodic point set, we present a fingerprint consisting of density functions that facilitates the efficient search for new materials and material properties. We prove invariance under isometries, continuity, and completeness in the generic case, which are necessary features for the reliable comparison of crystals. The proof of continuity integrates methods from discrete geometry and lattice theory, while the proof of generic completeness combines techniques from geometry with analysis. The fingerprint has a fast algorithm based on Brillouin zones and related inclusion-exclusion formulae. We have implemented the algorithm and describe its application to crystal structure prediction.

## 1 Introduction

This paper considers a deceptively simple question: *given periodic point sets (crystals) in  $\mathbb{R}^3$ , determine how close the sets are to being isometric*. In other words, how much do the points need to be perturbed to allow for a rigid transformation between the two sets? More generally, we may ask for the organization of a collection of periodic sets that facilitates efficient search. In this context, a *periodic (point) set* is the Minkowski sum of a *lattice* and a *motif*. The lattice is spanned by three linearly independent vectors, and the motif is a finite set of points in the *unit cell*, which is the parallelepiped whose edges are translates of the three vectors. The main reason for the difficulty of the question is the complicated nature of the continuous space of isometry classes of periodic sets:

- There is no method for choosing a unique basis for a lattice in a continuous manner. Indeed, continuity contradicts uniqueness as we can continuously deform a basis to a different basis of the same lattice. For example, the Niggli reduced cell [15] is unique but not continuous with respect to perturbations of the lattice [1].
- Crystallographers often use the symmetry group of crystals, which define a stratification of the space of isometry classes. Belonging to a given stratum is however not a continuous property.
- Small perturbations of a periodic set can significantly change the lattice with respect to which it is periodic.

Periodic sets are usually given by a basis of a lattice and the motif within the unit cell spanned by this basis. As explained above, even for extremely similar periodic sets, their lattices, their bases, and thus their motifs can look completely different, making a direct comparison between the motifs impossible. It is therefore advisable not to compute the

distances between the isometry classes but instead map the sets to a less complicated metric space. This is the approach we take in this paper. Part of the challenge is to determine which properties this map should possess, and how to balance its mathematical properties with efficient computability. As in many applications, we consider false positives in the comparison of two crystals less problematic than false negatives: a large distance in the metric space should imply that the two periodic sets are far from isometric, while a small distance should indicate a high chance that the sets are indeed close to being isometric.

The main contribution of this paper is a candidate solution, which we refer to as the *density fingerprint map*. For different non-negative integers,  $k$ , and for different non-negative radii,  $t$ , it maps the periodic set to the probability that a random point in the unit cell is at distance at most  $t$  from exactly  $k$  points in the periodic set; see Definition 1.

**Main Theorem.** *The density fingerprint maps a periodic set in  $\mathbb{R}^3$  to a series of density functions that satisfy the following properties:*

1. *the map is invariant under isometries (rigid motions and reflections) of space;*
2. *the map is Lipschitz continuous with respect to small perturbations of the points, with the Lipschitz constant depending on the packing and covering radii of the periodic set;*
3. *the map is generically complete: for a dense and open subset of the space of periodic sets, the isometry class of the periodic set is uniquely determined.*

Short of proving completeness beyond the generic case, we leave the completeness of the density fingerprint map for all periodic sets as an open question. Indeed, the authors of this paper failed to produce a counterexample to general completeness, but not because of a lack of trying. The Main Theorem generalizes to arbitrary finite dimensions.

The quest for a fingerprint map is motivated by the study of crystals, for which the configuration of atoms is important for their chemical properties. Quantifying the similarity between crystals has the potential to greatly improve the practice of *Crystal Structure Prediction*, which has come to rely on high-performance computing for simulating millions of structures. Many of them are similar to each other, and very few are eventually produced in the laboratory. An effective notion of similarity would allow an improved organization of the structures and lead to vast savings of supercomputing time currently wasted on redundant simulations. The prior work in this area is best summarized by listing the software systems currently used in practice: COMPACT [5], MERCURY [13], and COMPSTRU [10]. These systems are of great help in comparing crystals, but they employ heuristics like cut-offs and tolerances, which come with the usual drawbacks. It is our ambition to develop the mathematical and computational foundations needed to overcome the current deficiencies.

**Outline.** Section 2 provides the necessary notation and terminology for lattices, periodic sets, and isometries. Section 3 introduces the density functions for a periodic set and the corresponding density fingerprint map. Section 4 proves that the density fingerprint map is continuous with respect to perturbations of the periodic set. Section 5 proves that the density fingerprint map is complete for generic periodic sets. Section 6 explains how the density fingerprint is computed using the Brillouin zones of the points. Section 7 describes a preliminary application of the density fingerprint map to Crystal Structure Prediction. Section 8 concludes the paper.

## 2 Background

We cover two topics: locally finite point sets modeling crystals and transformations between them.

### 2.1 Delone Sets and Periodic Sets

We recall that  $A \subseteq \mathbb{R}^3$  is *locally finite* if any compact subset of  $\mathbb{R}^3$  contains only finitely many points of  $A$ . It is a *Delone set* [6] if there exist  $r, R > 0$  such that every open ball of radius  $r$  contains at most one point of  $A$  and every closed ball of radius  $R$  contains at least one point of  $A$ . In other words, no two points of  $A$  can be closer than  $2r$  and no point of  $\mathbb{R}^3$  can be further from  $A$  than  $R$ . We refer to the largest such  $r$  as the *packing radius* and the smallest such  $R$  as the *covering radius* of  $A$ . A Delone set is necessarily infinitely large and its points are in a sense evenly spread out over the entire Euclidean space.

We get an important subclass starting with three linearly independent vectors,  $v_1, v_2, v_3 \in \mathbb{R}^3$ , which we call a *basis*. The set of integer combinations is the *lattice*,  $\Lambda$ , and the set of real combinations with coefficients in  $[0, 1)$  is the *unit cell*,  $U$ , the vectors span:

$$\Lambda = \{n_1 v_1 + n_2 v_2 + n_3 v_3 \mid n_1, n_2, n_3 \in \mathbb{Z}\}, \quad (1)$$

$$U = \{r_1 v_1 + r_2 v_2 + r_3 v_3 \mid 0 \leq r_1, r_2, r_3 < 1\}. \quad (2)$$

We call any finite set  $M \subseteq U$  a *motif* and  $M + \Lambda = \{x + v \mid x \in M, v \in \Lambda\}$  a *periodic (point) set*. By construction,  $M + \Lambda + v = M + \Lambda$  for every  $v \in \Lambda$ , and if this exhausts all translations that keep  $M + \Lambda$  invariant, then  $U$  is a *primitive unit cell* of  $M + \Lambda$ . Since the International Union of Crystallography (IUCr) allows lattices that are spanned by fewer than three linearly independent vectors, it calls lattices as defined above *full*. We observe that  $M + \Lambda$  is a Delone set if  $\Lambda$  is full and not if  $\Lambda$  is not full, and therefore we will be exclusively interested in full lattices and so will assume as much without mentioning the term.

It is important to keep in mind that the basis and therefore the primitive unit cell are not unique. This is illustrated in Figure 1, which shows three of the infinitely many bases of the hexagonal lattice:  $a$  together with  $b - a$ ,  $b$ , or  $b + a$ . Applying Niggli's algorithm for the *Niggli reduced cell* [15] to this particular lattice, there is an ambiguity between the bases  $\{a, b\}$  and  $\{a, b - a\}$ , because the projections of  $b$  and  $b - a$  onto the line of  $a$  both have length  $\frac{1}{2}\|a\|$ . The tie can be broken by preferring  $b$ , but this causes a discontinuity in the construction of the reduced unit cell.

### 2.2 Rigid Motions and Isometries

Rather than a fixed set in  $\mathbb{R}^3$ , we often consider the class of sets that are equivalent under a particular type of transformation. For example, a *rigid motion* is a map  $\mathbb{R}^3 \rightarrow \mathbb{R}^3$  that is composed of a rotation and a translation. It preserves distances between pairs of points as well as orientations of ordered triplets of points. An *isometry* is a rigid motion possibly composed with a reflection, and so preserves distances but not necessarily orientations. This is the most relevant group of transformations to this paper as we model crystals by isometry classes of periodic sets.

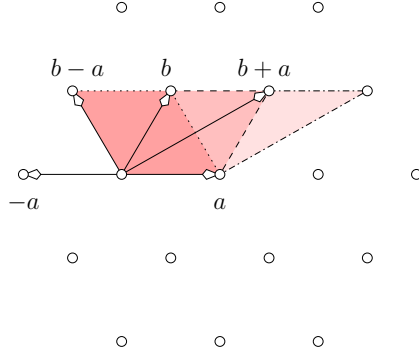


Figure 1: The bases  $\{a, b-a\}$ ,  $\{a, b\}$ ,  $\{a, b+a\}$  of the hexagonal lattice, and the corresponding unit cells drawn as shaded parallelograms.

### 3 The Density Fingerprint and its Invariance

A continuous invariant for comparing crystals is the density, defined as the total volume of balls centered at points in the motif divided by the volume of the unit cell. To avoid the choice of radii, we grow the balls continuously and simultaneously from their centers and get a 1-dimensional function rather than a single number. There are still many periodic sets this function cannot distinguish, for example any hexagonal close packing from the face-centered cubic lattice. We therefore add information by distinguishing points covered by a different number of balls.

**Definition 1** (Density Functions and Fingerprint). *Let  $A = M + \Lambda \subset \mathbb{R}^3$  be a periodic set and write  $A(t)$  for the collection of closed balls,  $B(a; t)$ , of radius  $t \geq 0$  centered at the points  $a \in A$ . The  $k$ -fold cover of  $A(t)$ , denoted  $\bigcup^k A(t)$ , consists of all points  $x \in \mathbb{R}^3$  contained in  $k$  or more of these balls. The fractional volume of the  $k$ -fold cover,  $\varphi_k^A(t) = \text{Vol}[U \cap \bigcup^k A(t)] / \text{Vol}[U]$ , is also the probability that a point chosen uniformly at random within a unit cell,  $U$ , belongs to at least  $k$  balls, and subtracting the fractional volume of the  $(k+1)$ -fold cover, we get the probability that the random point belongs to exactly  $k$  balls:*

$$\varphi_k^A(t) = \text{Prob}[x \in B(a; t) \text{ for } k \text{ or more points } a \in A]; \quad (3)$$

$$\psi_k^A(t) = \varphi_k^A(t) - \varphi_{k+1}^A(t) = \text{Prob}[x \in B(a; t) \text{ for exactly } k \text{ points } a \in A]. \quad (4)$$

We call  $\psi_k^A: [0, \infty) \rightarrow [0, 1]$  the  $k$ -th density function of  $A$ . The density fingerprint of  $A$  is the vector of density functions:  $\Psi(A) = (\psi_0^A, \psi_1^A, \dots, \psi_k^A, \dots)$ , and  $A \mapsto \Psi(A)$  is the density fingerprint map.

See Figure 2, which illustrates the density functions for the hexagonal and the square lattices in  $\mathbb{R}^2$ . Note that the density fingerprint is an isometry invariant and that it neither depends on the lattice used to write  $A$  as a periodic set, nor on its basis.

**Lemma 1** (Invariance under Isometries). *Let  $A \subseteq \mathbb{R}^3$  be a periodic set, and let  $Q \subseteq \mathbb{R}^3$  be isometric to  $A$ . Then  $\Psi(A) = \Psi(Q)$ .*

*Proof.* Let now  $\text{iso}: \mathbb{R}^3 \rightarrow \mathbb{R}^3$  be the isometry for which  $Q = \text{iso}(A)$ , and note that it also maps  $A(t)$  to  $Q(t)$  and  $\bigcup^k A(t)$  to  $\bigcup^k Q(t)$  for every  $k \geq 0$ . It follows that  $\psi_k^A(t) = \psi_k^Q(t)$ , for every  $k \geq 0$ , and therefore  $\Psi(A) = \Psi(Q)$ , as claimed.  $\square$

While the fingerprint map is not invariant under similarities, we can write  $\Psi(sA) = (\psi_0^A \circ s, \psi_1^A \circ s, \dots, \psi_k^A \circ s, \dots)$ , in which  $s(t) = st$  scales the radius. It would therefore be easy to construct a fingerprint map that is invariant under similarities, namely by normalizing the radius, e.g. by letting the radius be  $tr$ , in which  $r$  is the packing radius of  $A$ .

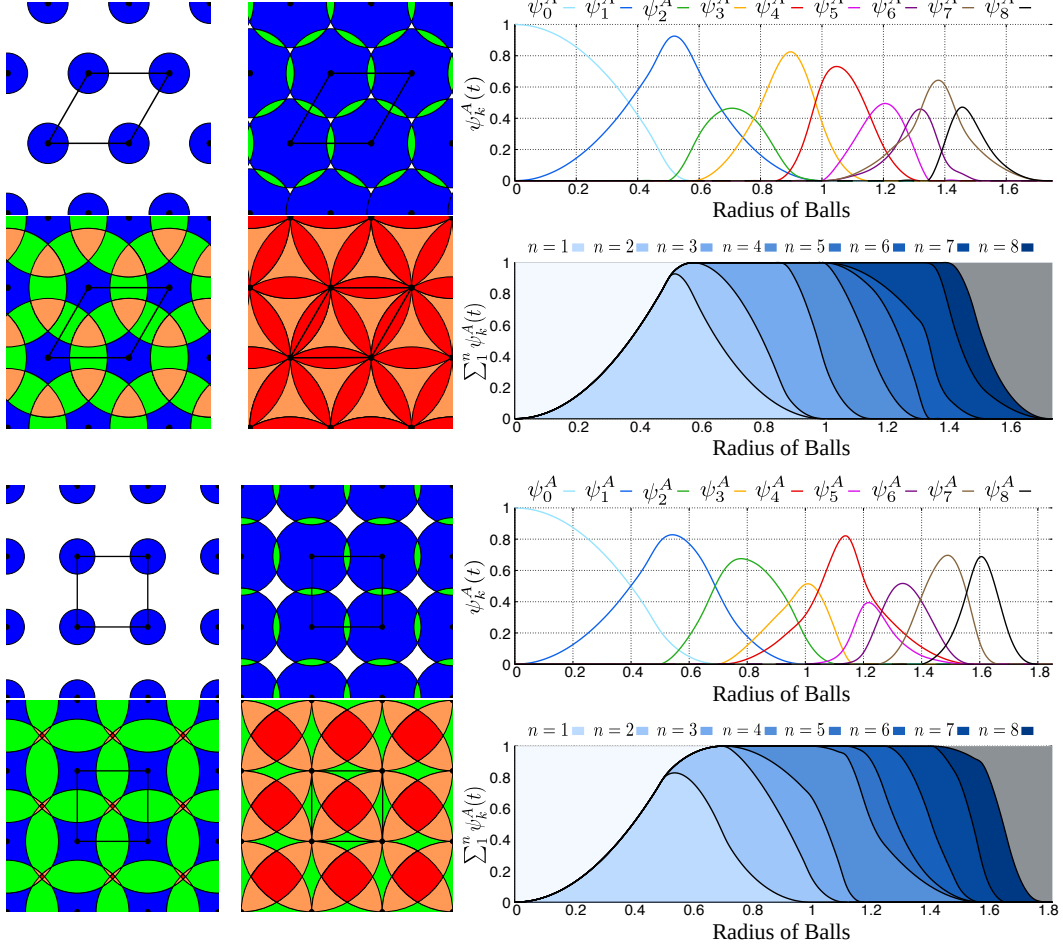


Figure 2: The density fingerprint map of the hexagonal lattice on the *top* and, for comparison, of the square lattice on the *bottom*. *Left*: the  $k$ -fold covers of the two sets for four different radii each:  $t = 0.25, 0.55, 0.75, 1.00$ . *Right*: the graphs of the respective first nine density functions above the corresponding *densigram*, in which the zeroth function can be seen upside-down and the remaining density functions are accumulated from left to right.

## 4 Continuity

We prove that the density fingerprint map is Lipschitz continuous with respect to small perturbations of the points. To formalize this result, we introduce distances between periodic sets and between density fingerprints. For sets  $A, Q \subseteq \mathbb{R}^3$  of equal cardinality, the (*Euclidean*) *bottleneck distance* is the infimum, over all bijections,  $\gamma: A \rightarrow Q$ , of the supremum

Euclidean distance between matched points, and for density fingerprints,  $\Psi(A), \Psi(Q)$ , we use the supremum of the weighted infinity norms of the differences between corresponding density functions:

$$d_B(A, Q) = \inf_{\gamma: A \rightarrow Q} \sup_{a \in A} \|a - \gamma(a)\|_2, \quad (5)$$

$$d_\infty(\Psi(A), \Psi(Q)) = \sup_{k \geq 0} \frac{1}{\sqrt[k+1]{k+1}} \|\psi_k^A - \psi_k^Q\|_\infty. \quad (6)$$

Note the damping of the difference between corresponding density functions. The reason for it is technical and related to the fact that density functions with higher  $k$  tend to vanish at later values of  $t$ . As a consequence, the sensitivity of the density function to any perturbation increases with growing  $k$ , and the damping compensates for this tendency. Before proving Lipschitz continuity, we show that two periodic sets with small bottleneck distance between them necessarily have a common lattice.

**Lemma 2** (Common Lattice). *Let  $A, Q$  be periodic sets in  $\mathbb{R}^3$ , and let  $r_Q > 0$  be the packing radius of  $Q$ . If  $d_B(A, Q) < r_Q$ , then there is a lattice  $\Lambda$  with unit cell  $U$  in  $\mathbb{R}^3$  such that  $\#(A \cap U) = \#(Q \cap U)$  and  $A = (A \cap U) + \Lambda$  and  $Q = (Q \cap U) + \Lambda$ .*

*Proof.* Since  $A, Q \subseteq \mathbb{R}^3$  are periodic, there are lattices with unit cells such that  $A = (A \cap U_A) + \Lambda_A$  and  $Q = (Q \cap U_Q) + \Lambda_Q$ . To get a contradiction, we assume that there is however no common lattice for  $A$  and  $Q$ . Equivalently,  $\Lambda_A \cap \Lambda_Q$  is a lattice of dimension at most 2. Therefore there exists a basis vector,  $v$ , of  $\Lambda_A$  such that  $nv \in \Lambda_Q$  implies  $n = 0$ . Picking a point  $a \in A$ , we consider the infinitely many points  $a(n) = a + nv$ , with  $n \in \mathbb{Z}$ . For each  $a(n)$ , let  $q(n)$  be the point in  $\Lambda_Q$  such that  $a(n) \in q(n) + U_Q$ , and define  $b(n) = a(n) - q(n)$ , which we note belongs to  $U_Q$ .

There are infinitely many pairwise different points  $b(n)$  in the unit cell, and it suffices to prove that at least one is at distance larger than  $\delta = d_B(A, Q)$  from all points in  $Q$ . To see this, let  $b(i)$  and  $b(j)$  be at distance less than  $\varepsilon = r_Q - \delta$  from each other, and note that  $b(i + n[j - i]) = b(i) + n[b(j) - b(i)]$ , for  $n \in \mathbb{Z}$ , provided the point on the right-hand side of the equation belongs to  $U_Q$ . In other words, we have an entire line of points with distance less than  $\varepsilon$  between contiguous points. The gap between balls of radius  $\delta$  centered at the points in  $Q$  is at least  $2\varepsilon$ , which implies that at least one of the points on the line is outside all such balls. This contradicts the assumption that the bottleneck distance between  $A$  and  $Q$  is  $\delta = r_Q - \varepsilon$ . The existence of a common lattice of  $A$  and  $Q$  follows.  $\square$

The proof of Lipschitz continuity makes use of the common lattice of the sets before and after the perturbation. We therefore formulate the claim assuming that the bottleneck distance between the two sets is less than the packing radii.

**Theorem 1** (Fingerprint Continuity). *Let  $A, Q$  be periodic sets in  $\mathbb{R}^3$ , both with packing radius at least  $r > 0$  and with covering radius at most  $R < \infty$ . If  $\delta = d_B(A, Q) < r$ , then there exists a constant  $C = C(r, R)$  such that  $d_\infty(\Psi(A), \Psi(Q)) \leq C \cdot d_B(A, Q)$ .*

*Proof.* By Lemma 2, there is a lattice,  $\Lambda \subseteq \mathbb{R}^3$ , that is common to both sets,  $A$  and  $Q$ , and we write  $U$  for the corresponding unit cell. Let  $\gamma: A \rightarrow Q$  be a bijection such that  $d_B(A, Q)$  is the supremum Euclidean distance between corresponding points, let  $k$  be a non-negative integer, and let  $t$  be a positive real number. We need an upper bound for

$$\left| \psi_k^A(t) - \psi_k^Q(t) \right| = \frac{|\text{Vol}[A_t^k \cap U] - \text{Vol}[Q_t^k \cap U]|}{\text{Vol}[U]}, \quad (7)$$

in which  $A_t^k = \bigcup^k A(t) \setminus \bigcup^{k+1} A(t)$  consists of all points  $x \in \mathbb{R}^3$  contained in exactly  $k$  balls of  $A(t)$ , and similarly for  $Q_t^k$ . As a first step, we find an upper bound on the numerator,  $\Delta$ , for the case in which  $\gamma$  is the identity except for one point,  $a \in M$ , which it maps to  $q = \gamma(a) \in B(a; \delta)$ ; that is:  $Q = A \setminus (a + \Lambda) \cup (q + \Lambda)$ . A point  $x \in \mathbb{R}^3$  is possibly covered by a different number of balls before and after the perturbation only if  $x \in [B(a; t) \ominus B(q; t)] + \Lambda$ , with  $\ominus$  denoting the symmetric difference. Observe that this set is contained in  $[B(\frac{a+q}{2}; t + \frac{\delta}{2}) \setminus B(\frac{a+q}{2}; t - \frac{\delta}{2})] + \Lambda$ . Hence,

$$\Delta \leq \text{Vol}[B(a; t) \ominus B(q; t)] \leq \frac{4\pi}{3} \left[ \left(t + \frac{\delta}{2}\right)^3 - \left(t - \frac{\delta}{2}\right)^3 \right] = \frac{4\pi}{3} \left[ 3\delta t^2 + \frac{1}{4}\delta^3 \right]. \quad (8)$$

Perturbing one point of  $M$  after the other, we can bound the error by (8) each time. Using the intensity,  $\rho = \#M/\text{Vol}[U]$ , this implies

$$\left| \psi_k^A(t) - \psi_k^Q(t) \right| \leq \rho \Delta \leq \rho \frac{4\pi}{3} \left[ 3\delta t^2 + \frac{1}{4}\delta^3 \right]. \quad (9)$$

We can eliminate the dependence on  $t$  by observing that for each  $k$  there is a value of  $t$  beyond which the  $k$ -th density functions of  $A$  and  $Q$  vanish. To determine this value, consider a point  $y \in \mathbb{R}^3$  and the sets  $A \cap B(y; t)$  and  $Q \cap B(y; t)$ . By the definition of  $R$ , the balls of radius  $R$  centered at the points of  $A$  cover  $B(y; t - R)$ , and similarly for  $Q$ . It follows that the two sets contain at least  $(t/R - 1)^3$  points each. Setting  $k + 1 \leq (t/R - 1)^3$ , we see that for  $t \geq R\sqrt[3]{k+1} + R$ , both sets have at least  $k + 1$  points each. Equivalently,  $y$  is covered by at least  $k + 1$  balls of radius  $t$ . Since this holds for every point  $y \in \mathbb{R}^3$ , we have  $\psi_k^A(t) = \psi_k^Q(t) = 0$  for all  $t \geq R\sqrt[3]{k+1} + R$ . Note that  $R\sqrt[3]{k+1} + R \leq 2R\sqrt[3]{k+1}$  for all  $k \geq 0$ . Replacing  $t$  in (9) by the latter bound, we get

$$\frac{1}{\sqrt[3]{k+1}^2} \|\psi_k^A - \psi_k^Q\|_\infty \leq 16\pi\rho R^2\delta + \frac{\pi}{3}\rho\delta^3 \leq \frac{12R^2}{r^3}\delta + \frac{1}{4r^3}\delta^3, \quad (10)$$

in which we use  $\rho \frac{4\pi}{3} r^3 \leq 1$  to get the final inequality. Using  $\delta^2 < r^2 < R^2$ , this gives  $C = 13R^2/r^3$  as an upper bound for the Lipschitz constant.  $\square$

Figure 3 illustrates Theorem 1 for a periodic set,  $A$ , and its perturbation,  $Q$ , in  $\mathbb{R}^2$  by showing the first eight (undamped) density functions for both sets in different colors.

## 5 Completeness

The fingerprint map is *complete* if it is injective up to isometries; that is: non-isometric periodic sets are mapped to different fingerprints. We prove completeness generically, i.e. on a dense open subset; compare to [4, 18, 12]. The density fingerprint also distinguishes non-generic sets for which other means fail, as will be illustrated by an example in Section 5.2. The completeness of the fingerprint for all periodic sets, however, remains an open question. Indeed, at the time of writing this paper, the authors are not aware of a 3-dimensional counterexample to completeness, but there is a 1-dimensional counterexample due to Morteza Saghaian: letting  $U = \{0, 4, 9\}$  and  $V = \{0, 1, 3\}$ , it can be checked that the finite sets  $U + V$  and  $U - V$ , and the periodic sets  $15\mathbb{Z} + (U + V)$  and  $15\mathbb{Z} + (U - V)$  cannot be distinguished by the 1-dimensional density fingerprint map.

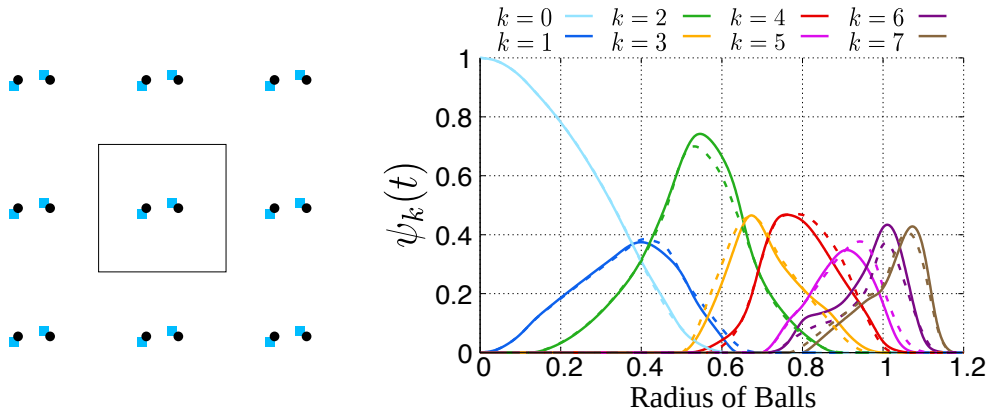


Figure 3: *Left*: a periodic set with two *black* points in its square unit cell, and the perturbed periodic set with two *blue* points in the same unit cell. *Right*: the graphs of the density functions are *solid* for the original set and *dashed* for the perturbed set. As predicted by Theorem 1, the small perturbation of the periodic set causes a small change in the fingerprint.

## 5.1 Generic Completeness

We prove the completeness of the density fingerprint map for generic periodic sets in  $\mathbb{R}^3$ . The notion of genericity is defined by conditions that are satisfied by an open and dense subset of the space of periodic point sets. We formulate such conditions in terms of the *circumradius* of edges, triangles, and tetrahedra, which is the radius of the smallest sphere that passes through the vertices of the simplex. To avoid infinitely many constraints, we introduce an upper bound on the circumradii to consider. Denoting by  $L = L(A, \vartheta)$  the list of all edges (pairs), triangles (triples) and tetrahedra (quadruples) spanned by points of a periodic set  $A$  whose circumradius is at most  $\vartheta$ , we call  $A$  *generic for a constant threshold*,  $\vartheta$ , if—apart from necessary violations due to periodicity—it satisfies the following three conditions:

1. the circumradii of different simplices in  $L$  are different;
2. the circumradii of different edges in  $L$  are not related to each other by a factor of 2.
3. for every  $t \leq \vartheta$ , there is at most one set of six circumradii of simplices in  $L$  such that the edges with twice their lengths assemble to a tetrahedron whose circumradius is  $t$ .

We call an edge a *lattice edge* if its length is the distance between two lattice points. Lattice edges violate Condition II and can thus be identified as such. A *lattice triangle* has three lattice edges, and a *lattice tetrahedron* has six lattice edges. The important difference between lattice and non-lattice simplices is that only the latter are unique up to lattice translations.

Since Conditions I, II, III can be phrased via finitely many algebraic equations in the vectors  $x \in M, v_1, v_2, v_3$ , the set of generic periodic sets with threshold  $\vartheta$  is open and dense (with respect to perturbations of these vectors) in the space of all periodic sets with at most  $m$  motif points. We write  $\text{Rad}(A)$  for the largest finite circumradius of  $p \leq 4$  points in  $A$  with pairwise distance at most four times the diameter of the unit cell. Since the diameter is the distance between two lattice points, this implies that  $\text{Rad}(A)$  is at least double the diameter.



**Theorem 2** (Generic Completeness). *Let  $A, Q \subseteq \mathbb{R}^3$  be non-isometric periodic sets that are generic for the threshold  $\vartheta = \max\{\text{Rad}(A), \text{Rad}(Q)\}$ . Then  $\Psi(A) \neq \Psi(Q)$ .*

*Proof.* Let  $[A]$  denote the isometry class of  $A$ . We prove the unique reconstruction of  $[A]$  from  $\Psi(A)$  in two steps:

$$\Psi(A) \xrightarrow{\text{STEP 1}} \text{tetrahedra in } L(A, \vartheta), \text{ up to isometries} \xrightarrow{\text{STEP 2}} [A]$$

STEP 1: Each density function is a weighted sum of the volumes of intersections of 2, 3, or 4 balls around points of  $A$ ; see [9, Equation (5)]. The volume formulas of such intersections are given in [8]. It is cumbersome but not difficult to prove that they are piecewise analytic, and that the circumradii of edges, triangles and tetrahedra spanned by points of  $A$  are the positions where the functions are not analytic. Therefore, the set of all positions up to  $\vartheta$  where at least one density function is not analytic yields the set of circumradii of simplices of  $L$ . We avoid the technicalities of using the differences between the left- and right-derivatives to distinguish which of these are caused by 2, 3, or 4 balls meeting, with the following trick. We treat all circumradii as if they were circumradii of edges, multiply them by two (to get the edge length), and try to assemble six of these edge lengths to form a tetrahedron. Whenever this gives a circumradius of a simplex of  $L$ , we have found a tetrahedron of  $A$  by Condition III. This way we can uniquely construct all tetrahedra of  $L$  up to isometries.

STEP 2: To start the process, we choose a non-lattice tetrahedron from the list. If there is no such tetrahedron, then  $A$  is a lattice and can be reconstructed from the lexicographically shortest lattice tetrahedron from the list—i.e. the tetrahedron consisting of the shortest lattice edge, the second-shortest lattice edge (linearly independent from the first), and so on—defining a (Minkowski-)reduced [14] and therefore primitive unit cell of  $A$ . On the other hand, if there exist non-lattice tetrahedra, we choose the lexicographically shortest one,  $abcd$ , with non-lattice edge  $ab$ .

Placing  $abcd$  in space—as we are only interested in the isometry class of  $A$ , we can place it arbitrarily—we identify all tetrahedra  $abce$  from the list that have  $abc$  as a face and try to glue them onto  $abcd$ . There are two possibilities (related by a reflection) of how to glue  $abce$ ; we denote the two different tip positions by  $e_1$  and  $e_2$ . We prove that at most one of the two options gives a positive result when checking if the tetrahedron  $abde_i$  is in the list of tetrahedra from Step 1: The triangles  $abd$  and  $abe$  are non-lattice, and therefore unique in  $A$  up to lattice translations by Condition I. Thus, when glued along  $ab$ , they span a uniquely defined tetrahedron  $abde$  with a certain edge length  $de$  that is the distance between  $d$  and  $e_i$  for at most one of its two possible positions.

This gluing procedure yields (among others) all points of distance at most four times the diameter of the unit cell to  $a, b, c, d$  (by definition of  $\text{Rad}(A)$ ), except the ones that lie on a plane spanned by the triangles  $abc$  or  $abd$ . This neighborhood is large enough such that it contains every motif point at least once and such that it contains a lattice basis, which can be identified by computing the pairwise differences between the reconstructed points and checking whether they satisfy Condition II. Repeating the reconstructed points with respect to the lattice yields the isometry class of  $A$ . As the construction was unique given the genericity conditions, we get  $\Psi(A) \neq \Psi(Q)$ .  $\square$

## 5.2 Distinguishing Non-Generic Periodic Sets

There are indications that the density fingerprint map distinguishes all periodic sets and not just the generic ones. We now give the reason for our optimism. Example 1 describes

two periodic sets that violate the above genericity conditions and can nevertheless be distinguished by the density fingerprint map. On the other hand, the two sets can neither be distinguished by their density nor by their X-ray diffraction patterns; two means commonly used in crystallography to determine the structure of a crystal. X-ray diffraction patterns give all pairwise distance vectors of the periodic set, but they do not determine the isometry class of a periodic set [16]: there exist *homometric structures*, which are non-isometric periodic sets with the same 2-point autocorrelation functions; that is: identical multisets of pairs, up to translation. There even exist periodic sets with the same 2- and 3-point autocorrelation functions, as we now explain.

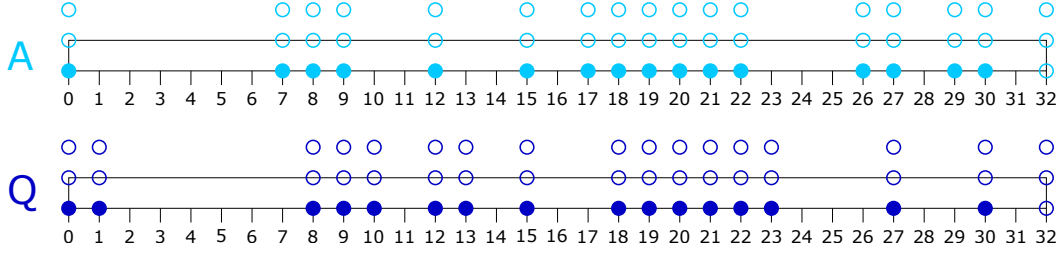


Figure 4: Periodic sets  $A$  and  $Q$  from Example 1, pictured with rectangular unit cells in two dimensions, for simplicity. Filled dots belong to the motifs while unfilled dots show the periodicity.

**Example 1.** Let  $A^{(1)}$  and  $Q^{(1)}$  be sets with periodicity 32 in  $\mathbb{R}$ , each with 16 points in the corresponding motif:

$$0, 7, 8, 9, 12, 15, 17, 18, 19, 20, 21, 22, 26, 27, 29, 30; \quad (11)$$

$$0, 1, 8, 9, 10, 12, 13, 15, 18, 19, 20, 21, 22, 23, 27, 30; \quad (12)$$

see Figure 4. The authors of [11, Section 5.3] show that  $A^{(1)}$  and  $Q^{(1)}$  have the same 2- and 3-point autocorrelation functions. Taking the Cartesian product with  $\mathbb{Z}^2$  preserves the equality between the autocorrelation functions, which yields periodic sets,  $A, Q \subseteq \mathbb{R}^3$ , with matching 2- and 3-point autocorrelation functions. Nevertheless, our density fingerprint map distinguishes them, as shown in Table 1: the  $L_\infty$ -distances between the first four corresponding density functions vanish but the next five  $L_\infty$ -distances are strictly positive.

$k$		0	1	2	3	4	5	6	7	8
$\ \psi_k^A - \psi_k^Q\ _\infty$		0.000	0.000	0.000	0.000	0.005	0.007	0.013	0.022	0.007

Table 1:  $L_\infty$ -distances between the corresponding density functions of the sets  $A$  and  $Q$  in Example 1.

## 6 Computation

The algorithm for the density fingerprint map is based on two related geometric concepts: the  $k$ -th Dirichlet–Voronoi domain and the  $k$ -Brillouin zone of a point. After introducing both, we explain how they are used, and how much time it takes to construct them.

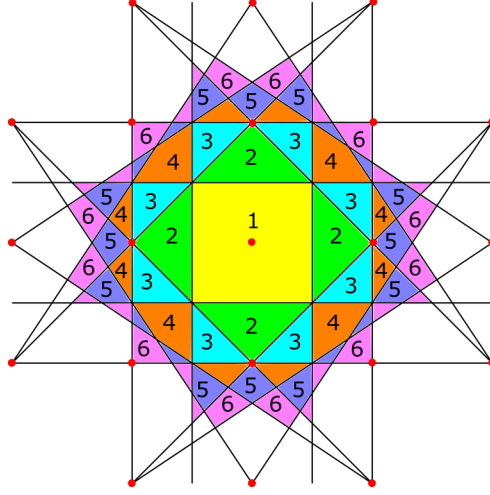


Figure 5: The 6-th Dirichlet–Voronoi domain of the point in the center decomposed into the first six Brillouin zones, which are indicated by colors and labels.

### 6.1 Dirichlet–Voronoi Domains and Brillouin Zones

Let  $A \subseteq \mathbb{R}^3$  be a locally finite set of points. For every positive integer  $k$ , the  $k$ -th *Dirichlet–Voronoi domain* of a point  $a \in A$  is the set of points in  $\mathbb{R}^3$  for which  $a$  is among the  $k$  closest points in  $A$ , and the  $k$ -th *Brillouin zone* is the difference between the  $k$ -th and the  $(k-1)$ -st Dirichlet–Voronoi domains:

$$\text{dom}_k(a, A) = \{x \in \mathbb{R}^3 \mid \|x - b\| < \|x - a\| \text{ for at most } k-1 \text{ points } b \in A\}, \quad (13)$$

$$\text{zone}_k(a, A) = \text{dom}_k(a, A) \setminus \text{dom}_{k-1}(a, A); \quad (14)$$

see Figure 5. Here we set  $\text{dom}_0(a, A) = \emptyset$  so that the first Brillouin zone is well defined. Note that  $\text{zone}_k(a, A)$  is the set of points  $x \in \mathbb{R}^3$  for which there are exactly  $k-1$  points  $b \in A$  that are closer to  $x$  than  $a$  is. Observe also that  $\text{dom}_k(a, A)$  is closed and star-convex, and if  $A$  is Delone, then it is also compact. If  $A$  is a lattice,  $A = \Lambda$ , then all  $k$ -th Dirichlet–Voronoi domains are translates of each other, and similarly for the Brillouin zones:  $\text{dom}_k(a, \Lambda) = \text{dom}_k(0, \Lambda) + a$  and  $\text{zone}_k(a, \Lambda) = \text{zone}_k(0, \Lambda) + a$ . Except for a measure zero subset of  $\mathbb{R}^3$ , every point  $x$  has a unique  $k$ -closest point in  $\Lambda$ . This implies that the  $k$ -th Brillouin zones *tile*  $\mathbb{R}^3$ , by which we mean that their closures cover  $\mathbb{R}^3$  while their interiors are pairwise disjoint. These properties generalize to a periodic set,  $A = M + \Lambda$ : the  $k$ -th Brillouin zones of the points in  $a + \Lambda$  are translates of each other, and the  $k$ -th Brillouin zones of all  $a \in A$  tile  $\mathbb{R}^3$ .

### 6.2 Decomposed Multiple Cover

Assume from here on that  $A = M + \Lambda$  is a periodic set. To compute  $\varphi_k^A$ , we may use any fundamental domain of the lattice. Particularly convenient is the union of the  $k$ -th Brillouin zones of the points in  $M$  as it lends itself to finding the subset covered by at least  $k$  of the balls.

**Theorem 3** (Density for Periodic Set). *Let  $A = M + \Lambda$  be periodic with lattice  $\Lambda \subseteq \mathbb{R}^3$  and motif  $M \subseteq U$  in the unit cell of  $\Lambda$ , and let  $k \geq 1$  be an integer. Then the probability that a*

random point  $x \in U$  belongs to  $k$  or more balls of radius  $t \geq 0$  centered at the points of  $A$  is

$$\varphi_k^A(t) = \frac{1}{\text{Vol}[U]} \sum_{a \in M} \text{Vol}[\text{zone}_k(a, A) \cap B(a; t)]. \quad (15)$$

*Proof.* Let  $M_k$  be the union of the  $k$ -th Brillouin zones of the points  $a \in M$  and note that  $M_k + \Lambda$  tiles  $\mathbb{R}^3$ . It follows that  $\text{Vol}[M_k] = \text{Vol}[U]$ . Let  $x \in M_k$  be in the interior of  $\text{zone}_k(a, A)$ . By construction,  $a$  is the unique  $k$ -closest point to  $x$ , so  $x$  lies in  $k$  or more balls if and only if  $x \in B(a; t)$ . Summing over all points  $a \in M$  gives (15).  $\square$

Clearly,  $\varphi_0^A(t) = 1$  for all radii  $t$ . Given  $k \geq 0$  and  $t \geq 0$ , we use (15) to compute  $\varphi_k^A(t)$  and  $\varphi_{k+1}^A(t)$ , and we get  $\psi_k^A(t) = \varphi_k^A(t) - \varphi_{k+1}^A(t)$ . To implement (15), we need to compute the volume of the intersection of a ball with a convex polyhedron. We could, for example, decompose the polyhedron into tetrahedra and use explicit expressions for the volume of intersections between balls and simplices; see for example [3]. A C++ implementation evaluating the density functions for a given periodic set using this strategy can be found at [19]. Alternatively, we could use inclusion-exclusion, which allows for further consolidation of the formula, writing  $\varphi_k^A(t)$  as an alternating sum of common intersections of up to four balls each. This does not lead to any asymptotic improvements of the running time, so we omit further details and refer to [9] instead.

### 6.3 Algorithm and Running Time

To evaluate the density functions  $\psi_0^A, \psi_1^A, \dots, \psi_k^A$  at a value  $t$ , we compute a plane arrangement for each point  $a \in M$  that consists of enough planes so that the first  $k+1$  Brillouin zones of  $a$  occur. Specifically, for a large enough radius,  $s$ , we consider for each  $b \neq a$  in  $A \cap B(a; s)$  the *bisector* of  $b$  and  $a$ , which is the plane defined by  $\|x - a\| = \|x - b\|$ . These bisectors decompose  $\mathbb{R}^3$  into convex cells. We refer to this decomposition as the *arrangement* of the planes. The 3-dimensional cells that are separated from  $a$  by exactly  $j-1$  planes form the  $j$ -th *belt* of the arrangement.

We now address the question how small we can choose  $s$  such that the first  $k+1$  belts are the first  $k+1$  Brillouin zones of  $a$ . To begin, we recall that  $t \geq 2R\sqrt[3]{k} + 1$  implies that  $\psi_i(t) = 0$  for  $0 \leq i \leq k$ ; see the proof of Theorem 1. To express this insight geometrically, let  $R_{k+1}(a)$  be the maximum distance of a point in the  $(k+1)$ -st Brillouin zone of  $a$  from  $a$ . That the density functions  $\psi_0^A$  to  $\psi_k^A$  are zero for  $t \geq 2R\sqrt[3]{k} + 1$  implies  $\varphi_{k+1}^A(t) = 1$ , for these values of  $t$ , and therefore  $R_{k+1}(a) \leq 2R\sqrt[3]{k} + 1$ . To capture all the relevant planes, it thus suffices to consider all points  $b \in A \setminus \{a\}$  at distance at most  $s = 2R_{k+1}(a)$  from  $a$ . Using a straightforward volume argument, we see that  $B(a; 2R_{k+1}(a))$  contains at most  $(4R\sqrt[3]{k} + 1 + r)^3 / r^3 = \mathcal{O}(k)$  points, in which we treat  $r$  and  $R$  as constants.

Constructing the arrangement of  $\mathcal{O}(k)$  planes incrementally, as described in [7, chapter 7], takes time  $\mathcal{O}(k^3)$ . Doing this for each point in the motif takes time  $\mathcal{O}(\#M \cdot k^3)$ , and within the same time bound we can evaluate the first  $k+1$  density functions.

## 7 An Application to Crystal Structure Prediction

Crystal Structure Prediction (CSP) aims to predict whether a selected molecule can exist as a functional material, i.e. a crystal with useful functions or properties. In other words, CSP seeks to answer the question of whether copies of a molecule can be arranged in such a way

that the resulting crystal is stable (will not deform and lose its properties over time) as well as useful. Crucially, CSP tries to answer this question without setting foot in a laboratory, with the hope of dramatically reducing the need to perform the time-consuming process of physically synthesizing crystals.

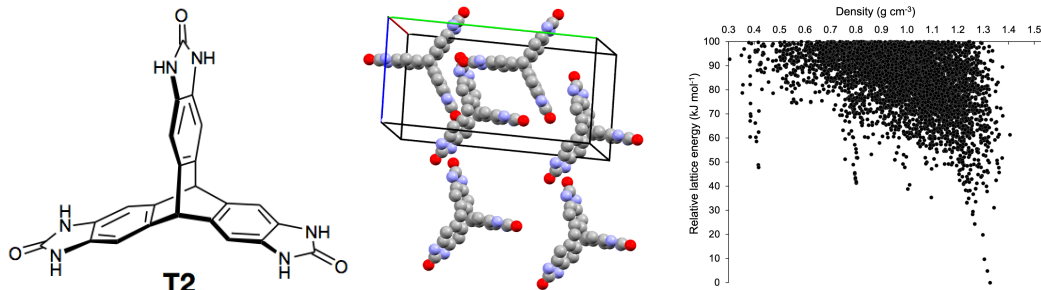


Figure 6: *Left*: a T2 molecule. *Middle*: the T2- $\delta$  crystal with highlighted unit cell. *Right*: the output of CSP for the T2 molecule. It is a plot of 5679 simulated T2 crystal structures [17, Fig. 2d], each represented by two coordinates: the physical density (atomic mass within a unit cell divided by the unit cell volume) and energy (determining the crystal’s thermodynamic stability). Structures at the bottom of the ‘downward spikes’ are likely to be stable.

Our collaborators at Liverpool’s Materials Innovation Factory [17] used CSP to predict that the T2 molecule (Figure 6) can be crystallized into a new structure that has half the physical density of the only previously known structure for T2, a desirable property for applications such as gas storage. As part of this process, they also identified four other structures of interest. Following the CSP predictions, they synthesized 5 families of T2-crystals in the laboratory by varying parameters like temperature and pressure, calling them T2- $\alpha$ , T2- $\beta$ , ..., T2- $\epsilon$ . One of them, T2- $\gamma$ , indeed had the desired property of having only half the physical density of the previously known structure T2- $\alpha$ . They scanned the synthesized crystals using X-ray powder diffraction yielding Crystallographic Information Files, each containing the unit cell and the motif points representing the atoms. These files were then compared with the results of the simulations, either by using their physical density alongside the COMPACT algorithm—which compares only a finite portion of the structure—or by looking at visualizations of the crystal structures. This comparison showed that the synthesized crystals matched the prediction well. Our collaborators deposited these structures into the globally used Cambridge Structural Database.

At a later time, we used our newly developed fingerprints to verify our collaborators’ matchings between the synthesized crystals T2- $\alpha$  to T2- $\epsilon$  and the simulated crystals entry 99, 28, 62, 09, 01. We did so by computing, for each of the five matches, the distance between the density functions of the synthesized and the simulated crystal. As one is the prediction of the other, we expected to see small distances. And for four of the five structures this was true: T2- $\gamma$ , for example, always has an  $L_\infty$ -distance of less than 0.04 over the first eight pairs of corresponding density functions; see Table 2. However, when we came to check the distances between density functions of T2- $\delta$  with its predicted structure, we were surprised to see large distances (the final row of Table 2). It turned out that a mix-up of files had happened, and what was uploaded to the Cambridge Structural Database as T2- $\delta$  was in fact T2- $\beta'$  (a crystal from the T2- $\beta$  family). The density fingerprint revealed this error,

which was verified by chemists upon a visual inspection, and it is because of this that T2- $\delta$  was subsequently correctly deposited.

$\ \psi_k^A - \psi_k^Q\ _\infty$	$k = 0$	1	2	3	4	5	6	7
T2- $\alpha$ vs entry 99	0.0042	0.0092	0.0125	0.0056	0.0099	0.0088	0.0127	0.0099
T2- $\beta$ vs entry 28	0.0157	0.0156	0.0159	0.0224	0.0334	0.0396	0.0357	0.0454
T2- $\gamma$ vs entry 62	0.0020	0.0080	0.0128	0.0155	0.0153	0.0250	0.0296	0.0391
T2- $\delta$ vs entry 09	0.0610	0.0884	0.1267	0.0676	0.0915	0.0801	0.0733	0.0388
T2- $\epsilon$ vs entry 01	0.0132	0.0152	0.0207	0.0571	0.0514	0.0431	0.0468	0.0550
T2- $\beta'$ vs entry 09	0.2981	0.2631	0.3718	0.3747	0.2563	0.2360	0.3161	0.3232

Table 2: *First five rows:* the  $L_\infty$ -distances between the first eight pairs of corresponding density functions of physically synthesized T2 crystals (T2- $\alpha$ , T2- $\beta$ , etc.) and the simulated structures that had predicted them from the CSP output dataset (entry XX). *Last row:* the suspiciously larger numbers revealed the mix-up of the files T2- $\delta$  and T2- $\beta'$  and thus led to depositing the initially omitted Crystallographic Information File of the T2- $\delta$  crystal into the Cambridge Structural Database.

Plots of the density functions of correctly matched synthesized and simulated structures can be seen in Figure 7. As another application, we expect that the fingerprint will be used to simplify the large output data sets produced by CSP by comparing simulated structures with each other, thus speeding up what is currently a slow process.

## 8 Discussion

The main contribution of this paper is a fingerprint map from periodic sets in  $\mathbb{R}^3$  (which model crystals) to series of density functions. This map is obviously invariant under isometries, and we prove it is continuous and generically complete. We leave the completeness without genericity assumption as an open question. In this context, it is worth noticing that our proof of generic completeness makes only limited use of the order,  $k$ , at which the circumradius of an edge, triangle, or tetrahedron is detected. Recall that the order is the number of points in the respective circumsphere. Is this additional information sufficient to prove general completeness?

A drawback of the bottleneck distance between periodic sets used in this paper is its sensitivity to changes of the unit cell; see Lemma 2. An alternative dissimilarity that may be more relevant in practice considers affine transformations,  $\tau$ , that minimize the bottleneck distance:

$$d_{\text{AT}}(A, Q) = \inf_{\tau} \max\{\min\{d_B(A, \tau(Q)), d_B(\tau(A), Q)\}, |\log s_1|, |\log s_3|\}, \quad (16)$$

in which  $s_1 \geq s_2 \geq s_3$  are the three singular values of the matrix of  $\tau$ . Is the density fingerprint map defined in Section 3 continuous with respect to this dissimilarity?

We close this paper with three extensions of the results presented in this paper. Different types of atoms are often modeled as balls with different radii. A possible geometric formalism is that of weighted points and the power distance [2]. Our geometric results generalize to this setting, but some need a careful adaptation. Our continuity result for periodic sets (Theorem 1) also generalizes to non-periodic Delone sets that allow for a reasonable

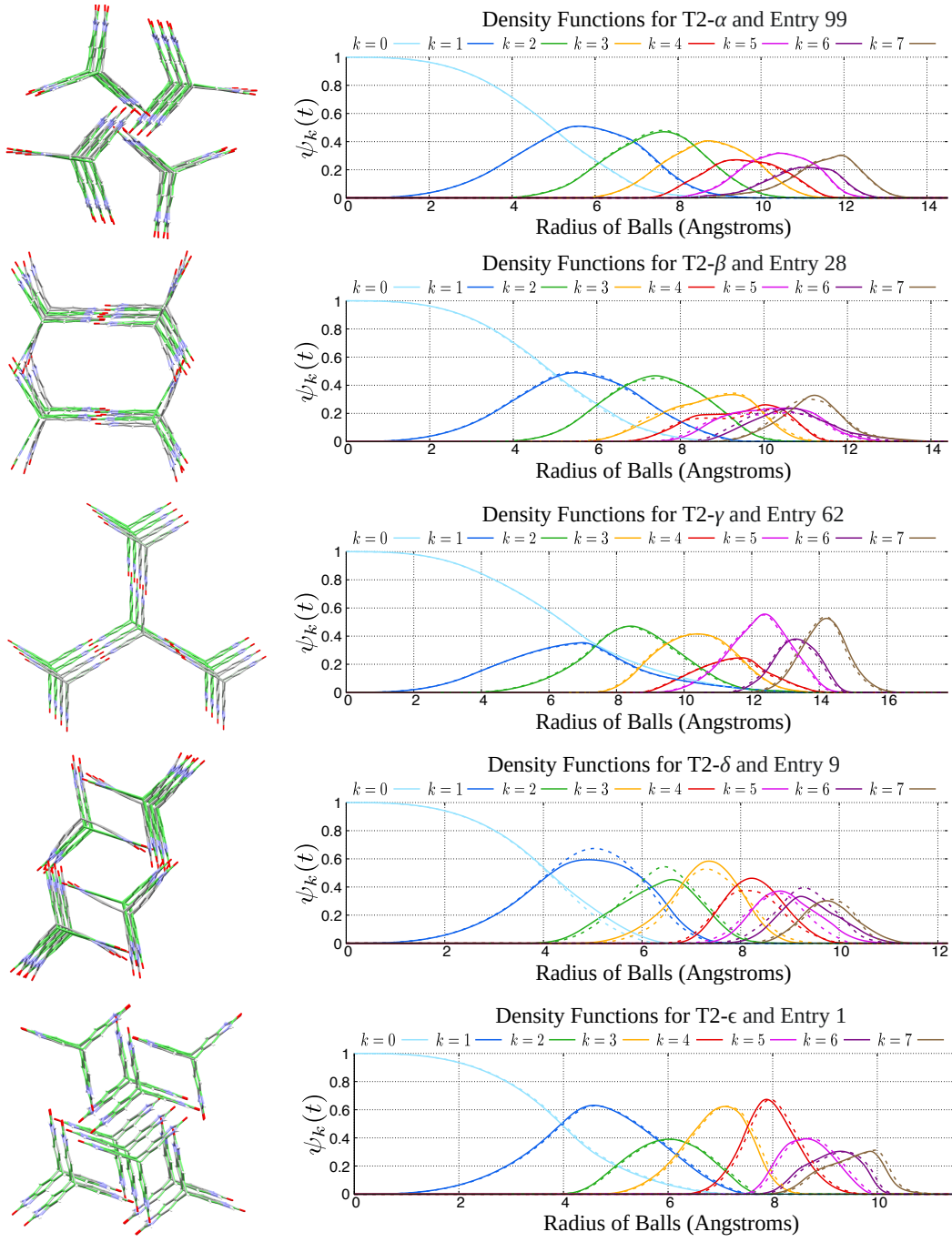


Figure 7: *Left*: experimental T2 crystals (curved gray molecules) and their simulated versions (straight green molecules) overlaid. *Right*: the density functions of the periodic sets of molecular centres of the experimental T2 crystals (solid curves) vs. simulated crystals (dashed curves).

definition of density functions. Considering that quasiperiodic crystals can be modeled as such, it might be worthwhile to find out how far such an extension can be pushed. Finally, we mention that our results generalize to arbitrary finite dimension.

## References

- [1] Lawrence C. Andrews, Herbert J. Bernstein, and G.A. Pelletier. “A perturbation stable cell comparison technique”. In: *Acta Crystallographica* A36.2 (1980), pp. 248–252. DOI: 10.1107/S0567739480000496.
- [2] Franz Aurenhammer. “Power diagrams: properties, algorithms and applications”. In: *SIAM Journal on Computing* 16.1 (1987), pp. 78–96. DOI: 10.1137/0216006.
- [3] David Avis, Binay K. Bhattacharya, and Hiroshi Imai. “Computing the volume of the union of spheres”. In: *The Visual Computer* 3.6 (1988), pp. 323–328. DOI: 10.1007/BF01901190.
- [4] Mireille Boutin and Gregor Kemper. “On reconstructing n-point configurations from the distribution of distances or areas”. In: *Advances in Applied Mathematics* 32.4 (2004), pp. 709–735. DOI: 10.1016/S0196-8858(03)00101-5.
- [5] James A. Chisholm and Sam Motherwell. “COMPACT: a program for identifying crystal structure similarity using distances”. In: *Journal of Applied Crystallography* 38.1 (2005), pp. 228–231. DOI: 10.1107/S0021889804027074.
- [6] Nikolai P. Dolbilin, Jeffrey C. Lagarias, and Marjorie Senechal. “Multiregular point systems”. In: *Discrete & Computational Geometry* 20.4 (1998), pp. 477–498. DOI: 10.1007/PL00009397.
- [7] Herbert Edelsbrunner. *Algorithms in Combinatorial Geometry*. Vol. 10. Springer-Verlag Berlin Heidelberg, 1987. DOI: 10.1007/978-3-642-61568-9.
- [8] Herbert Edelsbrunner and Ping Fu. “Measuring space filling diagrams and voids”. In: *Molecular Biophysics Report UIUC-BI-MB-94-01*. University of Illinois at Urbana-Champaign, 1994.
- [9] Herbert Edelsbrunner and Mabel Iglesias-Ham. “Multiple covers with balls I: inclusion–exclusion”. In: *Computational Geometry* 68 (2018), pp. 119–133. DOI: 10.1016/j.comgeo.2017.06.014.
- [10] Gemma de la Flor et al. “Comparison of structures applying the tools available at the Bilbao Crystallographic Server”. In: *Journal of Applied Crystallography* 49.2 (2016), pp. 653–664. DOI: 10.1107/S1600576716002569.
- [11] F. Alberto Grünbaum and Calvin C. Moore. “The use of higher-order invariants in the determination of generalized Patterson cyclotomic sets”. In: *Acta Crystallographica* A51.3 (1995), pp. 310–323. DOI: 10.1107/S0108767394009827.
- [12] Paul Lemke, Steven S. Skiena, and Warren D. Smith. *Reconstructing sets from inter-point distances*. Springer Berlin Heidelberg, 2003, pp. 597–631. DOI: 10.1007/978-3-642-55566-4\_27.
- [13] Clare F. Macrae et al. “Mercury: visualization and analysis of crystal structures”. In: *Journal of Applied Crystallography* 39.3 (2006), pp. 453–457. DOI: 10.1107/S002188980600731X.



- [14] Phong Q. Nguyen and Damien Stehlé. “Low-dimensional lattice basis reduction revisited”. In: *Proceedings of the 6th International Algorithmic Number Theory Symposium* (2004), pp. 338–357. DOI: 10.1007/978-3-540-24847-7\_26.
- [15] Paul Niggli. *Krystallographische und strukturtheoretische Grundbegriffe*. Vol. 1. Akademische verlagsgesellschaft mbH, 1928.
- [16] Linus Pauling and Maple D. Shappell. “8. The crystal structure of bixbyite and the C-modification of the sesquioxides”. In: *Zeitschrift für Kristallographie - Crystalline Materials* 75.1 (1930), pp. 128–142. DOI: 10.1515/zkri-1930-0109.
- [17] Angeles Pulido et al. “Functional materials discovery using energy–structure–function maps”. In: *Nature* 543.7647 (2017), pp. 657–664. DOI: 10.1038/nature21419.
- [18] Marjorie Senechal. “A point set puzzle revisited”. In: *European Journal of Combinatorics* 29.8 (2008), pp. 1933–1944. DOI: 10.1016/j.ejc.2008.01.013.
- [19] Philip Smith. *Density functions of a periodic set in C++*. 2020. URL: [https://github.com/Phil-Smith1/Density%5C\\_Functions%20\(2020\)](https://github.com/Phil-Smith1/Density%5C_Functions%20(2020)).