Supporting Information: Rapid prediction of molecular crystal structures using simple topological and physical descriptors

Nikolaos Galanakis[1*] and Mark E. Tuckerman[1,2,3,4*]

[1]Department of Chemistry, New York University, New York, 10003, New York, USA.
[2]Courant Institute of Mathematical Sciences, New York University, New York, 10012, New York, USA.
[3]NYU-ECNU Center for Computational Chemistry, NYU Shanghai, 3663 Zhongshan Road North, Shanghai, 200062, China.
[4]Simons Center for Computational Physical Chemistry at New York UniversityNew York,10003,New York,USA.

*Corresponding author(s). E-mail(s): ng1807@nyu.edu; mark.tuckerman@nyu.edu;

1 Example solutions to Eq. (10) for ZOPs

1.1 Space group $P2_1/c$

Space group $P2_1/c$ has 4 atoms in the unit cell. The symmetry operations are

$$\begin{aligned} &1:(x,y,z), && 2:(-x,-y,-z),\\ &3:(x,1/2-y,1/2+z), && 4:(-x,1/2+y,1/2-z). \end{aligned}$$

Equation (5) is heavily dependent on the periodic boundary conditions (PBC) since the application or not of the PBC along a specific direction defines the position of the pseudoparticle located in the center of mass of the 4 atoms. Let $p_i \in [0,1]$, $i = x,y,z$ a coefficient which states in the PBC are applied along a specific direction ($p_i = 1$) or not ($p_i = 0$). For $p_y = 0$ we have two atoms at (r_1,θ_1,ϕ_1), (r_2,θ_2,ϕ_2) and another two at $(r_3,\theta_3,-\phi_1)$, $(r_4,\theta_4,-\phi_2)$, for $(p_y,p_z) = (1,0)$ we have two atoms at (r_1,θ_1,ϕ_1), (r_3,θ_3,ϕ_3) and another two at $(r_2, \theta_2, \pi{+}\phi_1)$, $(r_4, \theta_4, \pi{+}\phi_3)$ while for $(p_y, p_z) = (1, 1)$ we have two atoms at (r_1, θ_1, ϕ_1), (r_3, θ_3, ϕ_3) and another two at $(r_1, \pi{-}\theta_1, \pi{+}\phi_1)$, $(r_3, \pi{-}\theta_3, \pi{+}\phi_3)$. For these cases, eq. (5) can be written as follows:

$$\begin{aligned} &p_y = 0 && [R_{n\ell}(r_1)P_\ell^{-m}(\cos\theta_1) - R_{n\ell}(r_3)P_\ell^{-m}(\cos\theta_3)]\sin(m\phi_1)+ \\ &m < 0 && +[R_{n\ell}(r_2)P_\ell^{-m}(\cos\theta_2) - R_{n\ell}(r_4)P_\ell^{-m}(\cos\theta_4)]\sin(m\phi_2) = 0. \end{aligned} \tag{S.1}$$

$$\begin{aligned} &p_y = 0 && [R_{n\ell}(r_1)P_\ell^{m}(\cos\theta_1) + R_{n\ell}(r_3)P_\ell^{m}(\cos\theta_3)]\cos(m\phi_1)+ \\ &m \geq 0 && +[R_{n\ell}(r_2)P_\ell^{m}(\cos\theta_2) + R_{n\ell}(r_4)P_\ell^{m}(\cos\theta_4)]\cos(m\phi_2) = 0. \end{aligned} \tag{S.2}$$

$$\begin{aligned} &(p_y, p_z) = (1, 0) && [R_{n\ell}(r_1)P_\ell^{-m}(\cos\theta_1) + (-1)^m R_{n\ell}(r_2)P_\ell^{-m}(\cos\theta_2)]\sin(m\phi_1)+ \\ &m < 0 && +[R_{n\ell}(r_3)P_\ell^{-m}(\cos\theta_3) + (-1)^m R_{n\ell}(r_4)P_\ell^{-m}(\cos\theta_4)]\sin(m\phi_3) = 0. \end{aligned} \tag{S.3}$$

$$\begin{aligned} &(p_y, p_z) = (1, 0) && [R_{n\ell}(r_1)P_\ell^{m}(\cos\theta_1) + (-1)^m R_{n\ell}(r_2)P_\ell^{m}(\cos\theta_2)]\cos(m\phi_1)+ \\ &m \geq 0 && +[R_{n\ell}(r_3)P_\ell^{m}(\cos\theta_3 2) + (-1)^m R_{n\ell}(r_4)P_\ell^{m}(\cos\theta_4)]\cos(m\phi_2) = 0. \end{aligned} \tag{S.4}$$

$$\begin{aligned} &(p_y, p_z) = (1, 1) && [1 + (-1)^\ell]\Big[R_{n\ell}(r_1)P_\ell^{-m}(\cos\theta_1)\sin(m\phi_1)+ \\ &m < 0 && +R_{n\ell}(r_3)P_\ell^{-m}(\cos\theta_3)\sin(m\phi_3)\Big] = 0, \end{aligned} \tag{S.5}$$

$$\begin{aligned} &(p_y, p_z) = (1, 1) && [1 + (-1)^\ell]\,[R_{n\ell}(r_1)P_\ell^{m}(\cos\theta_1)\cos(m\phi_1)+ \\ &m \geq 0 && +R_{n\ell}(r_3)P_\ell^{m}(\cos\theta_3)\cos(m\phi_3)] = 0. \end{aligned} \tag{S.6}$$

These equations provide non-trivial solutions only for $(p_x, p_y) = (1, 1)$ and $\ell = \text{odd}$. The non-trivial solutions are:

1. For $x = 0.5p_x$ we get $\phi_1 = \pm\pi/2$ and $\phi_2 = \pm\pi/2$ (for $p_y = 0$) or $\phi_3 = \pm\pi/2$ (for $p_y = 1$). In this case, all order parameters with $m = -2k$, $k \in \mathbb{N}^*$ and $m = 2k + 1$, $k \in \mathbb{N}$ are zero.

2. For $y = 0$ we have two different cases:

   (a) When $(p_y, p_z) = (0, 0)$ we have $r_4 = r_1$, $r_3 = r_2$, $\theta_4 = \pi - \theta_1$, $\theta_3 = \pi - \theta_2$ and $\phi_2 = \pi - \phi_1$ and equations (S.1)-(S.2) are

   $$\begin{aligned} &[1 + (-1)^\ell][R_{n\ell}(r_1)P_\ell^{-m}(\cos\theta_1) - (-1)^{\ell+m}R_{n\ell}(r_2)P_\ell^{-m}(\cos\theta_2)]\sin(m\phi_1) = 0, && m < 0, \\ &[1 + (-1)^\ell][R_{n\ell}(r_1)P_\ell^{m}(\cos\theta_1) + (-1)^{\ell+m}R_{n\ell}(r_2)P_\ell^{m}(\cos\theta_2)]\cos(m\phi_1) = 0, && m \geq 0, \end{aligned}$$

   and all the ZOPs for $\ell = \text{odd}$ are zero.

   (b) When $(p_y, p_z) = (0, 1)$ we have $r_2 = r_1$, $r_4 = r_3$, $\theta_2 = \pi - \theta_1$, $\theta_4 = \pi - \theta_3$ and $\phi_2 = \pi - \phi_1$ and equations (S.1)-(S.2) become

   $$\begin{aligned} &[1 - (-1)^\ell][R_{n\ell}(r_1)P_\ell^{-m}(\cos\theta_1) - R_{n\ell}(r_3)P_\ell^{-m}(\cos\theta_3)]\sin(m\phi_1) = 0, && m < 0, \\ &[1 + (-1)^\ell][R_{n\ell}(r_1)P_\ell^{m}(\cos\theta_1) + R_{n\ell}(r_3)P_\ell^{m}(\cos\theta_3)]\cos(m\phi_1) = 0, && m \geq 0, \end{aligned}$$

   and all the ZOPs for $[\ell = \text{even}, m < 0]$ and $[\ell = \text{odd}, m \geq 0]$ are zero.

3. For $y = 0.25$ or $y = 0.75$ and $(p_y, p_z) = (1, 0)$ we have $r_4 = r_1$, $r_3 = r_2$, $\theta_4 = \pi - \theta_1$, $\theta_3 = \pi - \theta_2$ and $\phi_3 = \phi_1$ and equations (S.3)-(S.4) become

   $$[1 + (-1)^\ell][R_{n\ell}(r_1)P_\ell^{-m}(\cos\theta_1) + (-1)^m R_{n\ell}(r_2)P_\ell^{-m}(\cos\theta_2)]\sin(m\phi_1) = 0, \quad m < 0,$$

$$[1+(-1)^{\ell}][R_{n\ell}(r_1)P_\ell^m(\cos\theta_1)+(-1)^m R_{n\ell}(r_2)P_\ell^m(\cos\theta_2)]\cos(m\phi_1)=0, \qquad m\geq 0,$$

and so, all the ZOPs for $\ell =$ odd are zero.

4. For $z=0$ we have again two cases.

   (a) If $(p_y,p_z)=(0,0)$ we get $r_3=r_1$, $r_4=r_2$, $\theta_3=\pi-\theta_1$ and $\theta_4=\pi-\theta_2$. Consequently, equations (S.1)-(S.2) become

   $$[1-(-1)^{\ell+m}][R_{n\ell}(r_1)P_\ell^{-m}(\cos\theta_1)\sin(m\phi_1)+R_{n\ell}(r_2)P_\ell^{-m}(\cos\theta_2)\sin(m\phi_2)]=0, \quad m<0,$$
   $$[1+(-1)^{\ell+m}][R_{n\ell}(r_1)P_\ell^{m}(\cos\theta_1)\cos(m\phi_1)+R_{n\ell}(r_2)P_\ell^{m}(\cos\theta_2)\cos(m\phi_1)]=0, \quad m\geq 0,$$

   which are true for $[\ell+m=\text{even},\ m<0]$ and $[\ell+m=\text{odd},\ m\geq 0]$.

   (b) When $(p_y,p_z)=(1,0)$ we have $r_2=r_1$, $r_4=r_3$, $\theta_2=\theta_1$ and $\theta_4=\theta_3$. So, equations (S.3)-(S.4) become

   $$[1+(-1)^{m}][R_{n\ell}(r_1)P_\ell^{-m}(\cos\theta_1)\sin(m\phi_1)+R_{n\ell}(r_3)P_\ell^{-m}(\cos\theta_3)\sin(m\phi_3)]=0, \quad m<0,$$
   $$[1+(-1)^{m}][R_{n\ell}(r_1)P_\ell^{m}(\cos\theta_1)\cos(m\phi_1)+R_{n\ell}(r_3)P_\ell^{m}(\cos\theta_3)\cos(m\phi_1)]=0, \quad m\geq 0,$$

   which are true for every $m=2k+1$, $k\in\mathbb{Z}$.

5. For $z=0.25$ or $z=0.75$ and $(p_y,p_y)=(0,1)$ we have again $r_3=r_1$, $r_4=r_2$, $\theta_3=\pi-\theta_1$ and $\theta_4=\pi-\theta_2$. Consequently, equations (S.1)-(S.2) become

   $$[1-(-1)^{\ell+m}][R_{n\ell}(r_1)P_\ell^{-m}(\cos\theta_1)\sin(m\phi_1)+R_{n\ell}(r_2)P_\ell^{-m}(\cos\theta_2)\sin(m\phi_2)]=0, \quad m<0,$$
   $$[1+(-1)^{\ell+m}][R_{n\ell}(r_1)P_\ell^{m}(\cos\theta_1)\cos(m\phi_1)+R_{n\ell}(r_2)P_\ell^{m}(\cos\theta_2)\cos(m\phi_1)]=0, \quad m\geq 0,$$

   which are true for $[\ell+m=\text{even},\ m<0]$ and $[\ell+m=\text{odd},\ m\geq 0]$.

6. For $(x,y)=(0.25,0)$ and $(p_x,p_y)=(0,0)$ or $(p_x,p_y)=(1,0)$ we have $\phi_1=3\pi/2\pm\pi/4$ and $\phi_2=3\pi/2\mp\pi/4$. As a result, all order parameters with $m=-4k$, $k\in\mathbb{N}^*$ or $m=4k+2$, $k\in\mathbb{N}$ are zero.

7. For $(x,y)=(0.75,0)$ and $(p_x,p_y)=(1,0)$ we get $\phi_1=-\pi/4$ and $\phi_2=5\pi/4$. As a result, all order parameters with $m=-4k$, $k\in\mathbb{N}^*$ or $m=4k+2$, $k\in\mathbb{N}$ are zero.

8. For $(x,y)=(0.5,0.25)$ and $(p_x,p_y)=(0,0)$ we have $\phi_1=0$ and $\phi_2=5\pi/2\mp\pi/4$. As a result, all ZOPs with $m=-4k$, $k\in\mathbb{N}^*$ are zero.

9. For $(x,y)=(0.25,0.25)$ or $(x,y)=(0.25,0.75)$ and $p_y=1$ we have $\phi_1=(2k+1)\pi/4$ and $\phi_3=(2k+1)\pi/4$. As a result, all order parameters with $m=-4k$, $k\in\mathbb{N}^*$ or $m=4k+2$, $k\in\mathbb{N}$ are zero.

10. For $(x,y)=(0.75,0.25)$ or $(x,y)=(0.75,0.75)$ and $(p_x,p_y)=(1,1)$ we have $\phi_1=(2k+1)\pi/4$ and $\phi_3=(2k+1)\pi/4$ and all ZOPs with $m=-4k$, $k\in\mathbb{N}^*$ or $m=4k+2$, $k\in\mathbb{N}$ are zero.

11. When $(x,y)=(0.5,0)$ or $(x,y)=(0.5,0.5)$ and $(p_x,p_y)=(0,1)$ we get $(\phi_1,\phi_3)=(5\pi/4,0)$ and the ZOPs are zero for $m=-4k$, $k\in\mathbb{N}^*$.

12. When $(x, y) = (0, 0)$ or $(x, y) = (0, 0.5)$ and $(p_x, p_y) = (1, 1)$ we get $(\phi_1, \phi_3) = (0, 5\pi/4)$ or $(\phi_1, \phi_3) = (k\pi, 5\pi/4)$ and the ZOPs are zero for $m = -4k$, $k \in \mathbb{N}^*$.

13. For $(y, z) = (0, 0.5)$ or $(y, z) = (0.5, 0)$ and $(p_y, p_z) = (1, 1)$ we get $r_3 = r_1$, $(\theta_1, \phi_3) = (\pi/2, 0)$ or $(\theta_1, \phi_3) = (0, \pi/2)$ and we get zero order parameters for $[\ell = \text{even}, m = -2k + 1, k \in \mathbb{N}^*]$.

14. For $(y, z) = (0.25k_y, 0.25k_z)$, $k_x, k_y = 1, 3$ and $(p_y, p_z) = (1, 1)$ we get $r_3 = r_1$, $\theta_3 = \pi - \theta_1$ and $\phi_3 = \phi_1$. As a result, all order parameters for $[\ell = \text{even}, m = 2k + 1, k \in \mathbb{Z}]$ are zero.

2 Random structure generation

2.1 Molecular orientations using the inertia frame

Consider a reference molecule in a reference coordinate system in which the atomic coordinates are $\mathbf{r}_0 = (f_x, f_y, f_z)$. For generality, it is useful to use the body-fixed frame for the reference molecule, on which the inertia tensor is diagonal and the inertia eigenvectors coincides with the standard Cartesian vector basis $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ of the Cartesian coordinate system. To generate a random orientation for the molecule we start from the body-fixed and we apply a two step rotation: (1) rotate the body-fixed eigenvector $(0, 0, 1)$ onto the normal vector $\hat{k} = (k_x, k_y, k_z)$ and (2) rotate the molecule by an angle

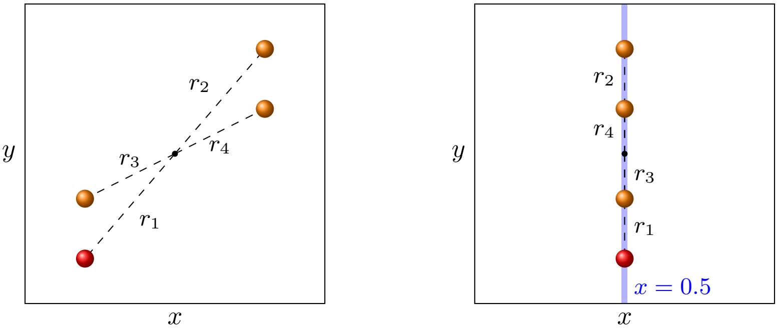


Fig. S.1 For a given reference atom (red) in a $P2_1/c$ space group, we can determine the positions of the symmetric atoms in the unit cell (orange) by simply applying the symmetry operations of the space group to the coordinates of the reference atom. When the reference atom is placed at a random position (Left), the ZOP's are in general non-zero, since each atom contributes different values in the sum of eq. (11). For certain positions of the reference atom, for example when $x = 0.5$ (right), the contributions between pairs of atoms for specific ℓ, m values are symmetric (in these case for the pairs 1-2 & 3-4) and the sum in eq. (11) becomes zero.

ω about $\hat{k}$. The rotation matrix for the first rotation is

$$R = \begin{pmatrix} 1 - Fk_x^2 & -Fk_xk_y & k_x \\ -Fk_xk_y & 1 - Fk_y^2 & k_y \\ -k_x & -k_y & k_z \end{pmatrix}, \qquad \text{where} \qquad F = \frac{1}{1+k_z} \tag{S.7}$$

and the combined rotation is given by

$$\mathbf{r} = (R\mathbf{r}_{0c})\cos\omega + \left[\hat{k} \times (R\mathbf{r}_0)\right]\sin\omega + \hat{k}\left[\hat{k} \cdot (R\mathbf{r}_0)\right](1 - \cos\omega) \tag{S.8}$$

Using this equation we can calculate the position of any position vector in the rotated system as

$$\begin{pmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \end{pmatrix} = \begin{pmatrix} e_{11} & e_{21} & k_x \\ e_{12} & e_{22} & k_y \\ e_{13} & e_{23} & k_z \end{pmatrix} \begin{pmatrix} f_x \\ f_y \\ f_z \end{pmatrix} \tag{S.9}$$

where

$$e_{11} = \cos\omega - Fk_x(k_x\cos\omega + k_y\sin\omega) \tag{S.10}$$
$$e_{12} = k_z\sin\omega - Fk_x(k_y\cos\omega - k_x\sin\omega) \tag{S.11}$$
$$e_{13} = -(k_x\cos\omega + k_y\sin\omega), \tag{S.12}$$
$$e_{21} = -k_z\sin\omega - Fk_y(k_x\cos\omega + k_y\sin\omega) \tag{S.13}$$
$$e_{22} = \cos\omega - Fk_y(k_y\cos\omega - k_x\sin\omega) \tag{S.14}$$
$$e_{23} = -(k_y\cos\omega - k_x\sin\omega) \tag{S.15}$$

Since the initial principal axes of inertia are the vectors of the standard basis, for a given rotation, it is easy to calculate the inertia eigenvectors in the rotated system

$$\hat{e}_1 = \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \end{pmatrix}, \qquad \hat{e}_2 = \begin{pmatrix} e_{21} \\ e_{22} \\ e_{23} \end{pmatrix}, \qquad \hat{e}_3 = \begin{pmatrix} e_{31} \\ e_{32} \\ e_{33} \end{pmatrix} = \begin{pmatrix} k_x \\ k_y \\ k_z \end{pmatrix} \tag{S.16}$$

where

$$e_{11}^2 + e_{12}^2 + e_{13}^2 = 1, \quad e_{21}^2 + e_{22}^2 + e_{23}^2 = 1, \quad e_{31}^2 + e_{32}^2 + e_{33}^2 = 1. \tag{S.17}$$

without knowing the details for the molecule. This way, the orientation of the molecule is defined by the principal axes of inertia in the rotated system, regardless the geometrical characteristics of the molecule.

In case where $\mathbf{r}_0 = (f_x, f_y, f_z)$ are the bond vectors of the atoms in the body-fixed coordinate system, we can use equation (S.9) to calculate the bond vectors in the rotated frame and then calculate the bond vectors in the fractional coordinate system using

$$\begin{pmatrix} \tilde{u} \\ \tilde{v} \\ \tilde{w} \end{pmatrix} = \mathbf{T} \begin{pmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \end{pmatrix} = \mathbf{T} \begin{pmatrix} e_{11} & e_{21} & e_{31} \\ e_{12} & e_{22} & e_{32} \\ e_{13} & e_{23} & e_{33} \end{pmatrix} \begin{pmatrix} f_x \\ f_y \\ f_z \end{pmatrix} \tag{S.18}$$

where

$$\mathbf{T} = \begin{pmatrix} \frac{1}{a} & -\frac{cos\gamma}{a\sin\gamma} & bc\frac{\cos\alpha\cos\gamma-\cos\beta}{\Omega\sin\gamma} \\ 0 & \frac{1}{b\sin\gamma} & ac\frac{\cos\beta\cos\gamma-\cos\alpha}{\Omega\sin\gamma} \\ 0 & 0 & \frac{ab\sin\gamma}{\Omega} \end{pmatrix} \tag{S.19}$$

is the transformation matrix from Cartesian to fractional coordinates in the case where a vector is aligned parallel to $\hat{x}$ axis and b vector is found on the Oxy plane. In this equation

$$\Omega = abc\sqrt{1-\cos^2\alpha-\cos^2\beta-\cos^2\gamma+2\cos\alpha\cos\beta\cos\gamma} \tag{S.20}$$

is the volume of the unit cell. While the Cartesian bond vectors are a function of the three variables describing the random rotation of the molecule, namely (k_x, k_y, k_z, ω) with $k_x^2 + k_y^2 + k_z^2 = 1$, the fractional bond vectors have an additional dependence on the 6 unit cell geometry variables $(a, b, c, \alpha, \beta, \gamma)$.

2.2 Molecular orientations and symmetry operations

Let $(\tilde{x}, \tilde{y}, \tilde{z})$ be the cartesian bond vectors for the atoms of the reference molecule. Consider a symmetric molecule in the unit cell, for which the atomic coordinates can be obtained by applying a symmetry operation on the atomic coordinates of the reference molecule. For the triclinic, monoclinic and orthorhombic crystal systems, the bond vectors for the second molecule will be given by

$$\begin{pmatrix} \tilde{x}_s \\ \tilde{y}_s \\ \tilde{z}_s \end{pmatrix} = \begin{pmatrix} p_a & 0 & 0 \\ 0 & p_b & 0 \\ 0 & 0 & p_c \end{pmatrix} \begin{pmatrix} \tilde{x} \\ \tilde{y} \\ \tilde{z} \end{pmatrix} \tag{S.21}$$

with $p_i \in \{-1, 1\}$. The bond vectors for the reference molecule can be calculated directly using eq. (S.9). The task is to find the value of angle ω_s, corresponding to the rotation of the symmetric molecule, so that the atomic coordinates of the atoms for the second molecule are consistent with (S.21). In general, the atomic coordinates for the atoms of the second molecule will be

$$\begin{pmatrix} \tilde{x}_s \\ \tilde{y}_s \\ \tilde{z}_s \end{pmatrix} = \begin{pmatrix} e_{11,s} & e_{21,s} & e_{31,s} \\ e_{12,s} & e_{22,s} & e_{32,s} \\ e_{13,s} & e_{23,s} & e_{33,s} \end{pmatrix} \begin{pmatrix} f_x \\ f_y \\ f_z \end{pmatrix} \tag{S.22}$$

where

$$\hat{e}_{s,1} = \begin{pmatrix} e_{s,11} \\ e_{s,12} \\ e_{s,13} \end{pmatrix}, \qquad \hat{e}_{s,2} = \begin{pmatrix} e_{s,21} \\ e_{s,22} \\ e_{s,23} \end{pmatrix}, \qquad \hat{e}_{s,3} = \begin{pmatrix} e_{s,31} \\ e_{s,32} \\ e_{s,33} \end{pmatrix} \tag{S.23}$$

are the inertia eigenvectors for the symmetric molecule. These eigenvectors must be consistent with the applied symmetry operations, and so we will have

$$\begin{pmatrix} e_{s,i1} \\ e_{s,i2} \\ e_{s,i3} \end{pmatrix} = \begin{pmatrix} p_a & 0 & 0 \\ 0 & p_b & 0 \\ 0 & 0 & p_c \end{pmatrix} \begin{pmatrix} e_{i1} \\ e_{i2} \\ e_{i3} \end{pmatrix} \tag{S.24}$$

These equations can be satisfied when

$$\sin\omega_s = \frac{p_c}{k_x^2 + k_y^2} \left[(p_b - p_a) k_x k_y \cos\omega + (p_a k_x^2 + p_b k_y^2) \sin\omega \right], \tag{S.25}$$

$$\cos\omega_s = \frac{p_c}{k_x^2 + k_y^2} \left[(p_a k_x^2 + p_b k_y^2) \cos\omega - (p_b - p_a) k_x k_y \sin\omega \right]. \tag{S.26}$$

The possible combinations for (p_a, p_b) are: $\{\pm(1,1), \pm(1,-1)\}$ in which cases we get

$$(p_a, p_b) = \pm(1,1): \qquad \sin\omega^{(2)} = \pm p_c \sin\omega, \tag{S.27}$$

$$\cos\omega_s = \pm p_c \cos\omega, \tag{S.28}$$

$$(p_a, p_b) = \pm(1,-1): \quad \sin\omega^{(2)} = \pm \frac{p_c}{p_a k_x^2 + p_b k_y^2} \left[2k_x k_y \cos\omega + (k_x^2 + k_y^2) \sin\omega \right], \tag{S.29}$$

$$\cos\omega_s = \pm \frac{p_c}{p_a k_x^2 + p_b k_y^2} \left[(k_x^2 + k_y^2) \cos\omega - 2k_x k_y \sin\omega \right]. \tag{S.30}$$

2.3 Molecular orientations and coplanarity

Let $Ax + By + Cz + D = 0$ be a plane in the reference unit cell, with $A^2 + B^2 + C^2 = 1$. Consider now a different unit cell K' of the structure. Let $\Delta\mathbf{r}_{uc,f} = (k_a, k_b, k_c)$ be the shift between the two unit cells in fractional coordinates. In Cartesian coordinates the shift will be

$$\Delta\mathbf{r}_{uc,c} = \begin{pmatrix} \Delta x \\ \Delta y \\ \Delta z \end{pmatrix} = \mathbf{H} \begin{pmatrix} k_a \\ k_b \\ k_c \end{pmatrix} \tag{S.31}$$

where

$$\mathbf{H} = \mathbf{T}^{-1} = \begin{pmatrix} a & b\cos\gamma & c\cos\beta \\ 0 & b\sin\gamma & c\frac{\cos\alpha - \cos\beta\cos\gamma}{\sin\gamma} \\ 0 & 0 & \frac{\Omega}{ab\sin\gamma} \end{pmatrix} \tag{S.32}$$

is the transformation matrix from fractional to Cartesian coordinates. The Cartesian equation of the image plane in unit cell K' will be

$$A(x + \Delta x) + B(y + \Delta y) + C(z + \Delta z) + D' = 0. \tag{S.33}$$

In order to preserve the coplanarity, the two planes must coincide, and this can happen only if $D = D'$, which yields

$$A\Delta x + B\Delta y + C\Delta z = 0. \tag{S.34}$$

or

$$\begin{aligned} &A(ak_a + bk_b\cos\gamma + ck_c\cos\beta)+ \\ &+ B\left(bk_b\sin\gamma + ck_c\frac{\cos\alpha - \cos\beta\cos\gamma}{\sin\gamma}\right) + C\frac{\Omega}{ab\sin\gamma}k_c = 0. \end{aligned} \tag{S.35}$$

This equation includes the 6 cell parameters $(a, b, c, \alpha, \beta, \gamma)$ and in the case where the plane $Ax + By + Cz + D = 0$ is related to the molecular orientation, the coefficients A, B, C are function of one of the three orientation angles of the molecule. Let $\hat{n}_p = (A, B, C)$ be the vector normal to the plane. For each vector $\hat{n}_p$ we can identify two linear independent crystallographic directions (k_a, k_b, k_c) that are perpendicular to $\hat{n}_p$ and as a result they preserve coplanarity. As a result, for a given plane vector $\hat{n}_p$, this equation can generate two constraints to the unit cell geometry and molecular orientation. If we are able to generate 9 constraints we will be able to generate a unique solution for the unit cell geometry and molecular orientation.

2.3.1 The coplanarity of the principal planes of inertia

Consider a molecule at a random orientation and let $\hat{e}_1$, $\hat{e}_2$, $\hat{e}_3$ be the eigenvectors of the inertia tensor corresponding to the principal axes of inertia. The general form of these vectors will be

$$\hat{e}_i = \begin{pmatrix} e_{i1} \\ e_{i2} \\ e_{i3} \end{pmatrix}, \qquad e_{i1}^2 + e_{i2}^2 + e_{i3}^2 = 1, \qquad i = 1, 2, 3. \tag{S.36}$$

In addition, the principal axes of inertia define a right-handed orthogonal coordinate system, and they must satisfy the equations

$$\hat{e}_i \cdot \hat{e}_j = 0, \qquad \hat{e}_i = \hat{e}_j \times \hat{e}_k, \qquad i \neq j \neq k, \qquad i \to j \to k \to i \tag{S.37}$$

Assuming that the inertia tensor eigenvectors preserve coplanarity along their normal planes (we will refer to these planes as principal inertia planes) they will satisfy equation

$$e_{i1}\Delta x + e_{i2}\Delta y + e_{i3}\Delta z = 0, \qquad\qquad i = 1, 2, 3. \tag{S.38}$$

Since there are infinite crystallographic directions in a crystal structure that can satisfy these conditions, it is important to identify a relatively small subset, that can be applied in every structure to get the correct molecular orientation. Based on the analysis of the $Z' = 1$ structures in the Cambridge Structural Database, a useful set of crystallographic directions is

$$\begin{aligned} &\mathbf{n}_{cd} = (n_u, n_v, n_w) \\ &\{n_u, n_v, n_w = 0, \pm 1, \ldots, \pm n_{\max}, \quad n_u|n_v|n_w = n_{\max}, \quad n_u n_v n_w = 0\}. \end{aligned} \tag{S.39}$$

with $n_{\max} \in \{1, 2, 3, 4, 5\}$. For each inertia eigenvector, eq. (S.38) can be applied twice, for two different crystallographic directions in the set $\mathbf{n}_{cd}$. This way, we get a set of 6 equations that can be used to implement constraints to the unit cell geometry and molecular orientation.

For a given molecular orientation, the inertia eigenvectors are uniquely defined, and the above 6 equations form a system with 6 unknown variables, the cell parameters $a, b, c, \alpha, \beta, \gamma$, which can be solved analytically to provide a unique cell geometry. In order to achieve that we need to know beforehand the crystallographic directions that preserve the coplanarity for the principle inertia planes which is not possible.

However, there are some constraints that can be applied to the possible crystallographic directions, to limit significantly the number of possible combinations. Consider the pairs of crystallographic directions $\{\mathbf{u}_{i1} = (u_{i11}, u_{i12}, u_{i13}), \mathbf{u}_{i2} = (u_{i21}, u_{i22}, u_{i23})$, perpendicular to the eigenvectors $\hat{e}_i$. Each pair, defines a crystallographic plane described by the vector

$$\mathbf{w}_i = (w_{i1}, w_{i2}, w_{i3}) = \tag{S.40}$$

$$= (u_{i12}u_{i23} - u_{i13}u_{i22}, u_{i13}u_{i21} - u_{i11}u_{i23}, u_{i11}u_{i22} - u_{i12}u_{i21}) \tag{S.41}$$

The cartesian form of the vector $\mathbf{w}_i$ which is perpendicular to the cartesian form of the plane defined by $\mathbf{u}_{i1} \times \mathbf{u}_{i2}$ and is given by

$$\mathbf{g}_i = \Omega \mathbf{T}^T \begin{pmatrix} w_{i1} \\ w_{i2} \\ w_{i3} \end{pmatrix} \tag{S.42}$$

and must be parallel to $\hat{e}_i$.

As a result, the vectors $\mathbf{e}_i$ must define an orthogonal coordinate system and they must be perpendicular to each other and so we get

$$\mathbf{g}_i \cdot \mathbf{g}_j = \Omega^2 (w_{i1}, w_{i2}, w_{i3}) \mathbf{T}\mathbf{T}^T \begin{pmatrix} w_{j1} \\ w_{j2} \\ w_{j3} \end{pmatrix} = 0 \tag{S.43}$$

The matrix product $\mathbf{T}\mathbf{T}^T$ is

$$\mathbf{T}\mathbf{T}^T = \begin{pmatrix} T_{11}^2 + T_{12}^2 + T_{13}^2 & T_{12}T_{22} + T_{13}T_{23} & T_{13}T_{33} \\ T_{12}T_{22} + T_{13}T_{23} & T_{22}^2 + T_{23}^2 & T_{23}T_{33} \\ T_{13}T_{33} & T_{23}T_{33} & T_{33}^2 \end{pmatrix} \tag{S.44}$$

where the individual components are given by

$$T_{11}^2 + T_{12}^2 + T_{13}^2 = \frac{b^2 c^2 \sin^2 \alpha}{\Omega^2} \tag{S.45}$$

$$T_{22}^2 + T_{23}^2 = \frac{a^2 c^2 \sin^2 \beta}{\Omega^2} \tag{S.46}$$

$$T_{33}^2 = \frac{a^2 b^2 \sin^2 \gamma}{\Omega^2} \tag{S.47}$$

$$T_{12}T_{22} + T_{13}T_{23} = \frac{abc^2 \sin\gamma(\cos\alpha\cos\beta - \cos\gamma)}{\Omega^2} \tag{S.48}$$

$$T_{13}T_{33} = \frac{ab^2 c \sin\gamma(\cos\alpha\cos\gamma - \cos\beta)}{\Omega^2} \tag{S.49}$$

$$T_{23}T_{33} = \frac{a^2 bc \sin\gamma(\cos\beta\cos\gamma - \cos\alpha)}{\Omega^2} \tag{S.50}$$

For monoclinic cells, eq. (S.43) becomes

$$(w_{i1}, w_{i2}, w_{i3}) \begin{pmatrix} b^2 c^2 & 0 & -ab^2 c \cos\beta \\ 0 & a^2 c^2 \sin^2\beta & 0 \\ -ab^2 c\cos\beta & 0 & a^2 b^2 \end{pmatrix} \begin{pmatrix} w_{j1} \\ w_{j2} \\ w_{j3} \end{pmatrix} = 0 \tag{S.51}$$

while for orthorhombic

$$(w_{i1}, w_{i2}, w_{i3}) \begin{pmatrix} b^2 c^2 & 0 & 0 \\ 0 & a^2 c^2 & 0 \\ 0 & 0 & a^2 b^2 \end{pmatrix} \begin{pmatrix} w_{j1} \\ w_{j2} \\ w_{j3} \end{pmatrix} = 0 \tag{S.52}$$

Equation (S.43) provide us with 3 additional constraints to the unit cell geometry that can be used as filters to accept or discard potential structures. By combining these with equations (S.38) we get a system of 9 equations that is able to be solved for every possible set of crystallographic directions that preserve the coplanarity of the principal inertia planes and provide us with a unique solution for the cell geometry and molecular orientation. The main advantage in this set of equations is that they do not depend explicitly on the molecular geometry. The information regarding the molecular orientation is included through the eigenvectors of the inertia tensor. In other words, by defining a set of crystallographic vectors that have high probability to be perpendicular to the eigenvectors of the inertia tensor, this set of equations has the ability to provide us with a set of general possible unit cell geometries and molecular orientations that can be valid for an given compound.

2.3.2 The coplanarity of the ring planes

The vectors $\hat{n}_r = (A_r, B_r, C_r)$ that are normal to the average planes of the atoms forming the rings in molecules, preserve coplanarity the same way as the principal inertia planes and as a result they will satisfy the equation

$$A_r \Delta x + B_r \Delta y + C_r \Delta z = 0. \tag{S.53}$$

Each of the rings in the molecule generates two additional constraints in the unit cell geometry and molecular orientation that can be used in combination with equations (S.38) and (S.43) to filter unreasonable structures. For a flat rigid molecule, e.g., coumarin, anthracene, pentacene,..., the plane of the ring(s) is identical to one of the principal axes of inertia, so the coplanarity of the rings does not provide additional information related to the geometry of the structure. For non-flat molecules however, like aspirin, we can use equations (S.38) and (S.43) to generate possible geometries and keep structures for which the normal ring vectors satisfy eq. (S.53) for at least two crystallographic directions in the set $\mathbf{n}_{cd}$. This equation can also be used for flexible molecules to determine the orientation of each rigid fragment in the reference molecule and generate possible conformers for a given unit cell.

2.3.3 Coplanar crystallographic directions: The set $\mathbf{n}_{cd}$

According to the CSD, for the vast majority of the organic molecular structures, each inertia eigenvector $\hat{e}$ is nearly perpendicular to at least two vectors $\mathbf{u}_{i1}, \mathbf{u}_{i2} \in \mathbf{n}_{cd}$. This set, for $n_{\max} \in \{1, 2, 3, 4, 5\}$, consists of 117 non-parallel vectors. Using pairs of vectors $\mathbf{u}_{i1}, \mathbf{u}_{i2} \in \mathbf{n}_{cd}$ we can generate 3097 possible non-parallel inertia eigenvectors $\mathbf{w}$. Let $\mathbf{W}^{(1)}$ be the set of the 3097 possible non-parallel inertia eigenvectors. Each triplet of vectors in the $\mathbf{W}^{(1)}$ has the potential to define a right-handed orthogonal system that represents the principal axes of inertia for a random molecule. Let $\mathbf{W}^{(3)}$ be the set of all possible triplets that can be generated using eigenvectors from the set $\mathbf{W}^{(1)}$. Without any constraints applied, there are $\frac{3097!}{3!} = 4\,945\,970\,940$ possible triplets in $\mathbf{W}^{(3)}$. However, the analysis showed that the angles between the eigenvectors in the crystallographic coordinate system satisfy specific conditions that allow us to reduce the number of triplets in the $\mathbf{W}^{(3)}$ to $1\,660\,859\,663$.

2.4 Atomic positions and the unit cell geometry

The Crystal Math approach is based on the fact that certain atoms in the unit cell (highly electropositive/electronegative atoms) are found near to the ZZPs in a way that determines the unit cell geometry. The general equation for the ZZP planes in crystallographic coordinates is

$$\epsilon_{\mathrm{ZZP}} : A_{\mathrm{ZZP}} u + B_{\mathrm{ZZP}} v + C_{\mathrm{ZZP}} w = 0.25 k_{\mathrm{ZZP}}, \tag{S.54}$$

where $A_{\mathrm{ZZP}}, B_{\mathrm{ZZP}}, C_{\mathrm{ZZP}} \in \{-1, 0, 1\}$, $A_{\mathrm{ZZP}} B_{\mathrm{ZZP}} C_{\mathrm{ZZP}} = 0$ and $k_{\mathrm{ZZP}} \in \{0, \pm 1, \pm 2, \pm 3, \pm 4\}$, while the distance between two parallel planes is

$$d^{\|}_{\mathrm{ZZP}} = \frac{0.25 |\delta k_{\mathrm{ZZP}}|}{\sqrt{A^2_{\mathrm{ZZP}} + B^2_{\mathrm{ZZP}} + C^2_{\mathrm{ZZP}}}}. \tag{S.55}$$

The exact values of the coefficients $A_{\mathrm{ZZP}}, B_{\mathrm{ZZP}}, C_{\mathrm{ZZP}}, k_{\mathrm{ZZP}}$ depend on the space group. The crystallographic coordinates for the atoms in the reference molecule are

$$(u, v, w) = (\tilde{u}, \tilde{v}, \tilde{w}) + (U, V, W) \tag{S.56}$$

Let atom i in the reference molecule be on a ZZP plane. The crystallographic coordinates for atom i must satisfy equation (S.54)

$$A_{\mathrm{ZZP}} \tilde{u}_i + B_{\mathrm{ZZP}} \tilde{v}_i + C_{\mathrm{ZZP}} \tilde{w}_i + A_{\mathrm{ZZP}} U + B_{\mathrm{ZZP}} V + C_{\mathrm{ZZP}} W = 0.25 k_{\mathrm{ZZP}} \tag{S.57}$$

or in more detail

$$\begin{aligned} A_{\mathrm{ZZP}} \sum_{i=1}^{3} T_{1i}(e_{1i} f_x + e_{2i} f_y + e_{3i} f_z) + B_{\mathrm{ZZP}} \sum_{i=2}^{3} T_{2i}(e_{1i} f_x + e_{2i} f_y + e_{3i} f_z) + \\ + C_{\mathrm{ZZP}} T_{33}(e_{13} f_x + e_{23} f_y + e_{33} f_z) + A_{\mathrm{ZZP}} U + B_{\mathrm{ZZP}} V + C_{\mathrm{ZZP}} W = 0.25 k_{\mathrm{ZZP}} \end{aligned} \tag{S.58}$$

2.4.1 Atomic separations and the unit cell geometry

Apart from having highly electropositive/electronegative atoms near the ZZPs, it is common to have pairs of atoms at separations $d^{\|}_{\mathrm{ZZP}}$ at directions perpendicular to the ZZP planes. Consider two atoms i and j in the reference molecule that are found at a distance $d^{\|}_{\mathrm{ZZP}}$ along a direction, perpendicular to the same ZZP plane. We will have that

$$A_{\mathrm{ZZP}}(\tilde{u}_i - \tilde{u}_j) + B_{\mathrm{ZZP}}(\tilde{v}_i - \tilde{v}_j) + C_{\mathrm{ZZP}}(\tilde{w}_i - \tilde{w}_j) = 0.25 \delta k_{\mathrm{ZZP}} \tag{S.59}$$

This can be written as

$$\begin{pmatrix} A_{\mathrm{ZZP}} & B_{\mathrm{ZZP}} & C_{\mathrm{ZZP}} \end{pmatrix} \mathbf{T} \begin{pmatrix} e_{11} & e_{21} & e_{31} \\ e_{12} & e_{22} & e_{32} \\ e_{13} & e_{23} & e_{33} \end{pmatrix} \begin{pmatrix} f_{x,i} - f_{x,j} \\ f_{y,i} - f_{y,j} \\ f_{z,i} - f_{z,j} \end{pmatrix} = 0.25 \delta k_{\mathrm{ZZP}} \tag{S.60}$$

For each pair of atoms that are found at a distance $d^{\|}_{\mathrm{ZZP}}$, the above equation generates a constraint for the 9 variables that define the unit cell geometry and the molecular orientation, namely the 6 cell geometry variables $(a, b, c, \alpha, \beta, \gamma)$ and the four orientational variables (k_x, k_y, k_z, ω) subject to the constraint $k_x^2 + k_y^2 + k_z^2 = 1$. For a triclinic cell, if we are able to identify 9 such atomic pairs, each along a direction perpendicular to a different ZZP, it is possible to determine a specific geometry for the unit cell. For monoclinic cells we need 7 atomic pairs while for orthorhombic unit cells we need just 6. To get the exact form of equations need to be solved, we can re-write equation (S.60) as

$$A_{\mathrm{ZZP}} \sum_{i=1}^{3} T_{1i}(e_{1i}\delta f_x + e_{2i}\delta f_y + e_{3i}\delta f_z) + B_{\mathrm{ZZP}} \sum_{i=2}^{3} T_{2i}(e_{1i}\delta f_x + e_{2i}\delta f_y + e_{3i}\delta f_z) + \\ + C_{\mathrm{ZZP}} T_{33}(e_{13}\delta f_x + e_{23}\delta f_y + e_{33}\delta f_z) = 0.25 k_{\mathrm{ZZP}}, \tag{S.61}$$

where T_{ij} are the non-zero components of the coordinate transformation matrix $\mathbf{T}$ and

$$\delta f_x = f_{x,i} - f_{x,j}, \qquad \delta f_y = f_{y,i} - f_{y,j}, \qquad \delta f_z = f_{z,i} - f_{z,j}.$$

For the separation along the different ZZP planes we get:

$$\epsilon_{\mathrm{ZZP}} : u = 0.25 k_{\mathrm{ZZP}} \\ \sum_{i=1}^{3} T_{1i}(e_{1i}\delta f_x + e_{2i}\delta f_y + e_{3i}\delta f_z) = 0.25 \delta k_{\mathrm{ZZP}} \tag{S.62}$$

$$\epsilon_{\mathrm{ZZP}} : v = 0.25 k_{\mathrm{ZZP}} \\ \sum_{i=2}^{3} T_{2i}(e_{1i}\delta f_x + e_{2i}\delta f_y + e_{3i}\delta f_z) = 0.25 \delta k_{\mathrm{ZZP}} \tag{S.63}$$

$$\epsilon_{\mathrm{ZZP}} : w = 0.25 k_{\mathrm{ZZP}} \\ T_{33} e_{13}\delta f_x + T_{33} e_{23}\delta f_y + T_{33} e_{33}\delta f_z = 0.25 \delta k_{\mathrm{ZZP}} \tag{S.64}$$

$$\epsilon_{\mathrm{ZZP}} : u \pm v = 0.25 k_{\mathrm{ZZP}} \\ \sum_{i=1}^{3} T_{1i}(e_{1i}\delta f_x + e_{2i}\delta f_y + e_{3i}\delta f_z) \pm \\ \pm \sum_{i=2}^{3} T_{2i}(e_{1i}\delta f_x + e_{2i}\delta f_y + e_{3i}\delta f_z) = 0.25 \delta k_{\mathrm{ZZP}} \tag{S.65}$$

$$\begin{aligned}
&\epsilon_{\mathrm{ZZP}} : u \pm w = 0.25 k_{\mathrm{ZZP}} \\
&\sum_{i=1}^{3} T_{1i}(e_{1i}\delta f_x + e_{2i}\delta f_y + e_{3i}\delta f_z) \pm \\
\pm &C_{\mathrm{ZZP}} T_{33}(e_{13}\delta f_x + e_{23}\delta f_y + e_{33}\delta f_z) = 0.25\delta k_{\mathrm{ZZP}}
\end{aligned} \tag{S.66}$$

$$\begin{aligned}
&\epsilon_{\mathrm{ZZP}} : v \pm w = 0.25 k_{\mathrm{ZZP}} \\
&\sum_{i=2}^{3} T_{2i}(e_{1i}\delta f_x + e_{2i}\delta f_y + e_{3i}\delta f_z) \pm \\
&\pm T_{33}(e_{13}\delta f_x + e_{23}\delta f_y + e_{33}\delta f_z) = 0.25\delta k_{\mathrm{ZZP}}
\end{aligned} \tag{S.67}$$

The above equations are expressed in terms of the components T_{ij} of the coordinate transformation matrix $\mathbf{T}$ and the components e_{ij} of the molecular rotation matrix. By substituting T_{ij} and e_{ij}, the above 9 equations generate a system with 9 unknown variables $(a, b, c, \alpha, \beta, \gamma, k_x, k_y, k_z, \omega)$ that if solved, it provides possible geometries for the unit cell.

Although equations (S.62)-(S.67) have the ability to generate a topologically sound set of random unit cell geometries, the number of possible atomic pairs makes the direct application of the method unfeasible. For a given molecule comprising of N atoms, there are $N(N-1)/2$ possible atomic pairs. This number is very big, even for small molecules. For example, in the aspirin molecule ($N = 21$) there are 210. This number is quite large and it generates a vast amount of possible structures. For a search in a triclinic system, we need to assign a pair to each of the 9 ZZP directions and so, the possible pair combinations is $[N(N-1)/2]^9$. In the cases of the aspirin molecule the possible combinations are 1.59×10^{19} and 7.94×10^{20} respectively. Even if we limit the number of possible pairs and consequently the number of possible combinations, by excluding the hydrogen atoms, (according to the database analysis the hydrogen atoms do not contribute to the geometry of the unit cell), the possible combinations for the aspirin molecule still remains to high: 4.60×10^{15} and 1.06×10^{17} respectively and as a result, we use equations (S.62)-(S.67) as filters to discard structures generated using equations (S.38) and (S.43) that describe the coplanarity of the principal planes of inertia as well as the coplanarity of the rings.

2.5 Close contacts

Let $(\tilde{x}, \tilde{y}, \tilde{z})$ be the cartesian bond vectors of the atoms for the reference molecule found at a random orientation. The bond vectors for a symmetric molecule in a triclinic, monoclinic or orthorhombic unit cell, will be given by (S.21). Let's assume that we want to create a close contact between atoms i in the reference molecule and atom j_s in the symmetric molecule. To generate the close contact, we first move the symmetric molecule so that atom j_s is found at the same position as atom i_r. The displacement vector will be

$$\delta\mathbf{r}_{j_s\to i} = (\tilde{x}_i - p_a\tilde{x}_j, \tilde{y}_i - p_b\tilde{y}_j, \tilde{z}_i - p_c\tilde{z}_j) \tag{S.68}$$

As a result, the positions of atoms in the symmetric molecule will now be $(x_s, y_s, z_s) = (p_a\tilde{x}, p_b\tilde{y}, p_c\tilde{z}) + \delta\mathbf{r}_{j_s\to i}$. The next step is to move the symmetric molecule again, so that atom j_s is found at a distance $d_{\text{vdW}}^{(i,j_s)} = R_{\text{vdW}}^i + R_{\text{vdW}}^j - d_{\text{overlap}}^{(i,j_s)}$ where $d_{\text{overlap}}^{(i,j_s)}$ is the overlapping distance between the vdW spheres of the two atoms (close contact strength). Let $\delta\mathbf{v}_{j_s\to i} = (dx_{j_s\to i}, dy_{j_s\to i}, dz_{j_s\to i})$ be the new shift so that

$$\left[dx_{j_s\to i}^2 + dy_{j_s\to i}^2 + dz_{j_s\to i}^2\right]^{1/2} = d_{\text{vdW}}^{(i,j_s)}. \tag{S.69}$$

The final shift of the symmetric molecule will be

$$\Delta\mathbf{r}_{j_s\to i} = \delta\mathbf{r}_{j_s\to i} + \delta\mathbf{v}_{j_s\to i} = (\tilde{x}_i - p_a\tilde{x}_j, \tilde{y}_i - p_b\tilde{y}_j, \tilde{z}_i - p_c\tilde{z}_j) + (dx_{j_s\to i}, dy_{j_s\to i}, dz_{j_s\to i}) \tag{S.70}$$

while the final positions of atoms in the symmetric molecule will be

$$\begin{aligned}(x_s, y_s, z_s) = (p_a\tilde{x}, p_b\tilde{y}, p_c\tilde{z}) + (\tilde{x}_i - p_a\tilde{x}_j, \tilde{y}_i - p_b\tilde{y}_j, \tilde{z}_i - p_c\tilde{z}_j) +& \\ (dx_{j_s\to i}, dy_{j_s\to i}, dz_{j_s\to i}) + (X, Y, Z),&\end{aligned} \tag{S.71}$$

where (X, Y, Z) is the position of the reference molecule.

2.5.1 Close contacts and the unit cell geometry

A pair of molecules forming a close contact can provide us with information regarding the geometry of the unit cell. The fractional coordinates for the atoms of the reference molecule are given by (S.56) while positions of the atoms in the symmetric molecule discussed in the previous paragraph will be

$$\begin{pmatrix} u_s \\ v_s \\ w_s \end{pmatrix} = \begin{pmatrix} p_a & 0 & 0 \\ 0 & p_b & 0 \\ 0 & 0 & p_c \end{pmatrix} \begin{pmatrix} u \\ v \\ w \end{pmatrix} + \begin{pmatrix} q_a \\ q_b \\ q_c \end{pmatrix} + \begin{pmatrix} n_a \\ n_b \\ n_c \end{pmatrix} \tag{S.72}$$

with $q_i \in \{0, 1/4, 1/2\}$[1] and $n_i \in \{-2, -1, 0, 1, 2\}$. The values n_i are used to generate the periodic image of the reference molecule in a neighbouring unit cell. The above equation can be written as

$$(u_s, v_s, w_s) = (p_a\tilde{u} + p_aU + q_a + n_a, p_b\tilde{v} + p_bV + q_b + n_b, p_c\tilde{w} + p_cW + q_c + n_c). \tag{S.73}$$

[1] For the triclinic, monoclinic and orthorhombic space groups, q_i get the value 1/4 only for spacegroups $Fdd2$ and $Fddd$

This equation in equivalent to the transformation of (S.71) from physical to fractional coordinates and so we get

$$\mathbf{T}\left(\begin{pmatrix} p_a\tilde{x} \\ p_b\tilde{y} \\ p_c\tilde{z} \end{pmatrix} + \begin{pmatrix} \tilde{x}_i \\ \tilde{y}_i \\ \tilde{z}_i \end{pmatrix} - \begin{pmatrix} p_a\tilde{x}_j \\ p_b\tilde{y}_j \\ p_c\tilde{z}_j \end{pmatrix} + \begin{pmatrix} dx_{j_s\to i} \\ dy_{j_s\to i} \\ dz_{j_s\to i} \end{pmatrix} + \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}\right) = \begin{pmatrix} p_a\tilde{u} \\ p_b\tilde{v} \\ p_c\tilde{w} \end{pmatrix} + \begin{pmatrix} p_aU \\ p_bV \\ p_cW \end{pmatrix} + \begin{pmatrix} q_a \\ q_b \\ q_c \end{pmatrix} + \begin{pmatrix} n_a \\ n_b \\ n_c \end{pmatrix} \tag{S.74}$$

Applying this equation for atom j we get

$$\mathbf{T}\begin{pmatrix} dx_{j_s\to i} \\ dy_{j_s\to i} \\ dz_{j_s\to i} \end{pmatrix} = \begin{pmatrix} p_a\tilde{u}_j - \tilde{u}_i + (p_a-1)U + q_a + n_a \\ p_b\tilde{v}_j - \tilde{v}_i + (p_b-1)V + q_b + n_b \\ p_c\tilde{w}_j - \tilde{w}_i + (p_c-1)W + q_c + n_c \end{pmatrix} \tag{S.75}$$

or

$$\begin{pmatrix} dx_{j_s\to i} \\ dy_{j_s\to i} \\ dz_{j_s\to i} \end{pmatrix} = \mathbf{H}\begin{pmatrix} p_a\tilde{u}_j - \tilde{u}_i + (p_a-1)U + q_a + n_a \\ p_b\tilde{v}_j - \tilde{v}_i + (p_b-1)V + q_b + n_b \\ p_c\tilde{w}_j - \tilde{w}_i + (p_c-1)W + q_c + n_c \end{pmatrix} \tag{S.76}$$

This equation provides a relation between the unit cell geometry $(a, b, c, \alpha, \beta, \gamma)$, the close contact and the position of the reference molecule. As a result, the creation of a close contact between a specific pair of atoms, generates constraints in the geometry of the unit cell. By assigning values to (U, V, W) so that the high charged atoms are found on the ZZPs of the space group under consideration and also to n_a, n_b, n_c it is possible to filter out structures generated using random inertia eigenvectors and ZZP equations that do not comply with equation (S.76).

Using the above equations, we can express the strength of a close contact as a function of the molecular position (U, V, W) in fractional coordinates and the parameters p_i, q_i, n_i according to equation

$$\begin{aligned} d_{\mathrm{vdW}}^{(i,j_s)} &= \left(dx_{j_s\to i}^2 + dy_{j_s\to i}^2 + dz_{j_s\to i}^2\right)^{1/2} \\ &= \big(a^2Q_1^2 + b^2Q_2^2 + c^2Q_3^2 + \\ &\quad + 2abQ_1Q_2\cos\gamma + 2acQ_1Q_3\cos\beta + 2bcQ_2Q_3\cos\alpha\big)^{1/2} \end{aligned} \tag{S.77}$$

where

$$Q_x^{(i,j_s)} = p_a\tilde{u}_j - \tilde{u}_i + (p_a-1)U + q_a + n_a \tag{S.78}$$

$$Q_y^{(i,j_s)} = p_b\tilde{v}_j - \tilde{v}_i + (p_b-1)V + q_b + n_b \tag{S.79}$$

$$Q_z^{(i,j_s)} = p_c\tilde{w}_j - \tilde{w}_i + (p_c-1)W + q_c + n_c \tag{S.80}$$

2.5.2 Hydrogen bond geometry

In the case of hydrogen bonds, the close contact is double: both the hydrogen atom and the electronegative atom (donor) that is covalently bound to the hydrogen, form close contacts with the highly electronegative hydrogen bond acceptor in the neighboring molecule. Let i be the hydrogen atom forming the hydrogen bond and j_s be the hydrogen bond acceptor in the neighboring molecule. The separation vector for the close contact in fractional coordinates is

$$\Delta \mathbf{r}^{(f)}_{j_s \to i} = Q_x \hat{u} + Q_y \hat{v} + Q_z \hat{w} \tag{S.81}$$

Let i' be the be the hydrogen bond donor in the reference molecule. The vector connecting atoms i and i' in fractional coordinates is

$$\Delta \mathbf{r}^{(f)}_{i' \to i} = (\tilde{u}_i - \tilde{u}_{i'})\hat{u} + (\tilde{v}_i - \tilde{v}_{i'})\hat{v} + (\tilde{w}_i - \tilde{w}_{i'})\hat{w} \tag{S.82}$$

In the vast majority of the hydrogen bonds, the vectors $\Delta \mathbf{r}^{(f)}_{j_s \to i}$ and $\Delta \mathbf{r}^{(f)}_{i' \to i}$ are almost parallel, a condition that can be express using equations

$$Q_y(\tilde{w}_i - \tilde{w}_{i'}) - Q_z(\tilde{v}_i - \tilde{v}_{i'}) = 0, \tag{S.83}$$
$$Q_x(\tilde{w}_i - \tilde{w}_{i'}) - Q_z(\tilde{u}_i - \tilde{u}_{i'}) = 0, \tag{S.84}$$
$$Q_x(\tilde{v}_i - \tilde{v}_{i'}) - Q_y(\tilde{u}_i - \tilde{u}_{i'}) = 0. \tag{S.85}$$

2.6 Conformer generation for flexible molecule search

Using the flexible molecule search, it is possible to generate difference conformers for a molecule. The cell geometries and molecular orientations are clustered based on the scaled unit cell geometry. In each cluster, there is a number of possible structures that differ only on the molecular orientation. By assigning the different molecular orientations to each fragment and joining the fragments to their common atom, we get different molecular conformations that are treated as rigid molecules in the subsequent steps of the protocol (Fig. S.2).

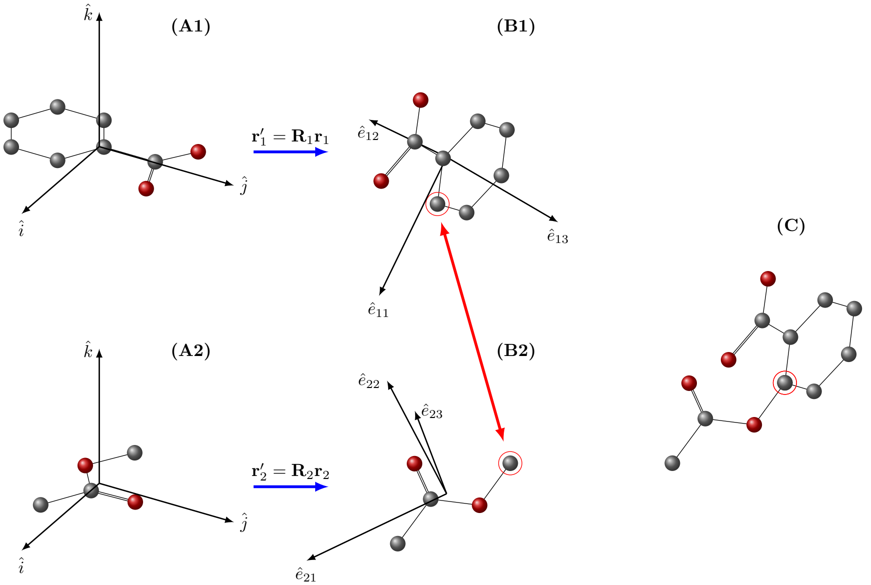


Fig. S.2 The conformation generation process in the case of aspirin molecule. In our search, aspirin molecule was treated as a flexible molecule composed of two rigid fragments. The fragments **(A1)** and **(A2)** are rotated using two different fragment orientation matrices $\mathbf{R}_1$, $\mathbf{R}_2$ corresponding to the same unit cell geometry. This rotations transforms the space fixed coordinates $\mathbf{r}_1$, $\mathbf{r}_2$ to the rotated coordinates $\mathbf{r}_1'$, $\mathbf{r}_2'$, shown in **(B1)** and **(B2)**. The complete molecule **(C)** is generated by joining the two fragments to their common atom, shown in red circle. By assigning various molecular orientations that correspond to the same unit cell geometries to each fragment, we can get different conformations for the molecule.

3 Cambridge Structural Database analysis

3.1 Molecular orientations

The statistical analysis of the $Z' \leq 5$ structures in the CSD database revealed that the principal axes of inertia as well as the vectors that are normal to rigid subgraphs of the molecules are almost perpendicular to at least one vector in

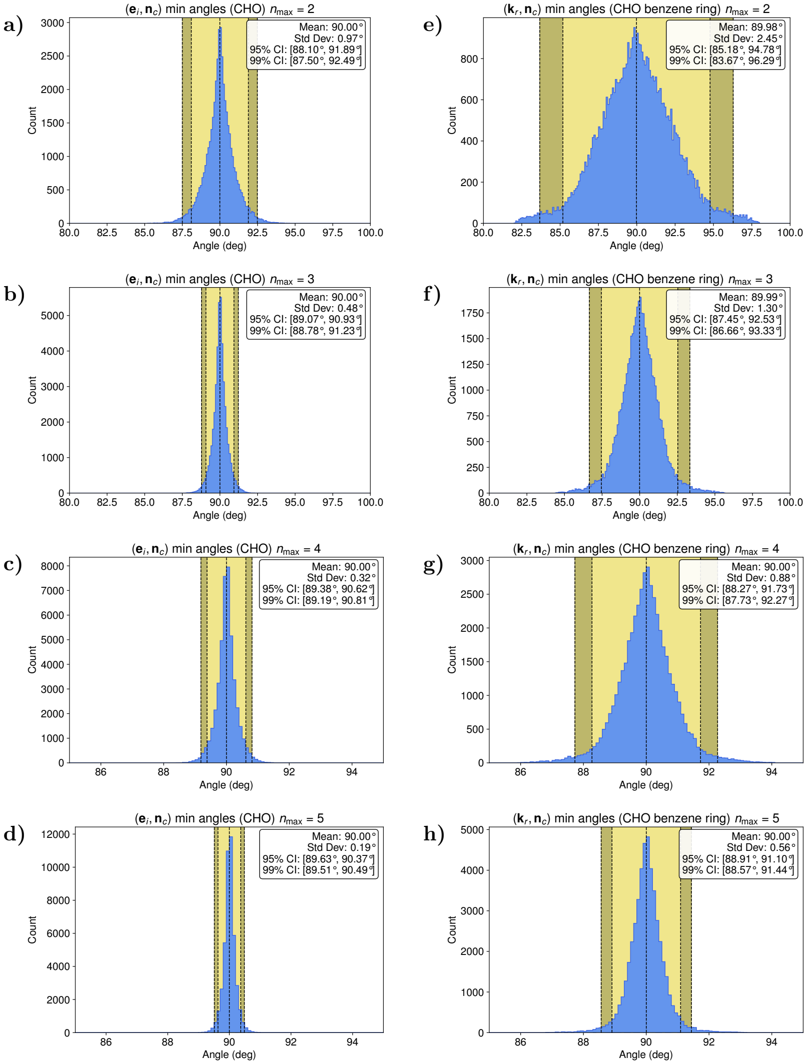


Fig. S.3 **a-d)** Distributions of the minimum angle formed by the vectors $\mathbf{e}_i$ and $\mathbf{n}_c$ for $n_{\max} = \{2, 3, 4, 5\}$ for all $Z' = 1$ structures composed of C, H, O atoms. **e-h)** Distributions of the minimum angle formed by the vectors $\mathbf{k}_r$ of the benzene rings and $\mathbf{n}_c$ for $n_{\max} = \{2, 3, 4, 5\}$ for all $Z' = 1$ structures composed of C, H, O atoms.

the set $\mathbf{n}_c$. In Figs. S.3, S.4 we show the distributions of the minimum angles formed by the vectors $(\mathbf{e}_i, \mathbf{n}_c)$ and $(\mathbf{k}_r, \mathbf{n}_c)$ for all structures composed of C, H, O and C, H, N, O atoms. The similarity between the distributions for the different molecular compositions suggest that the orientations are independent on the atomic species comprising the molecules.

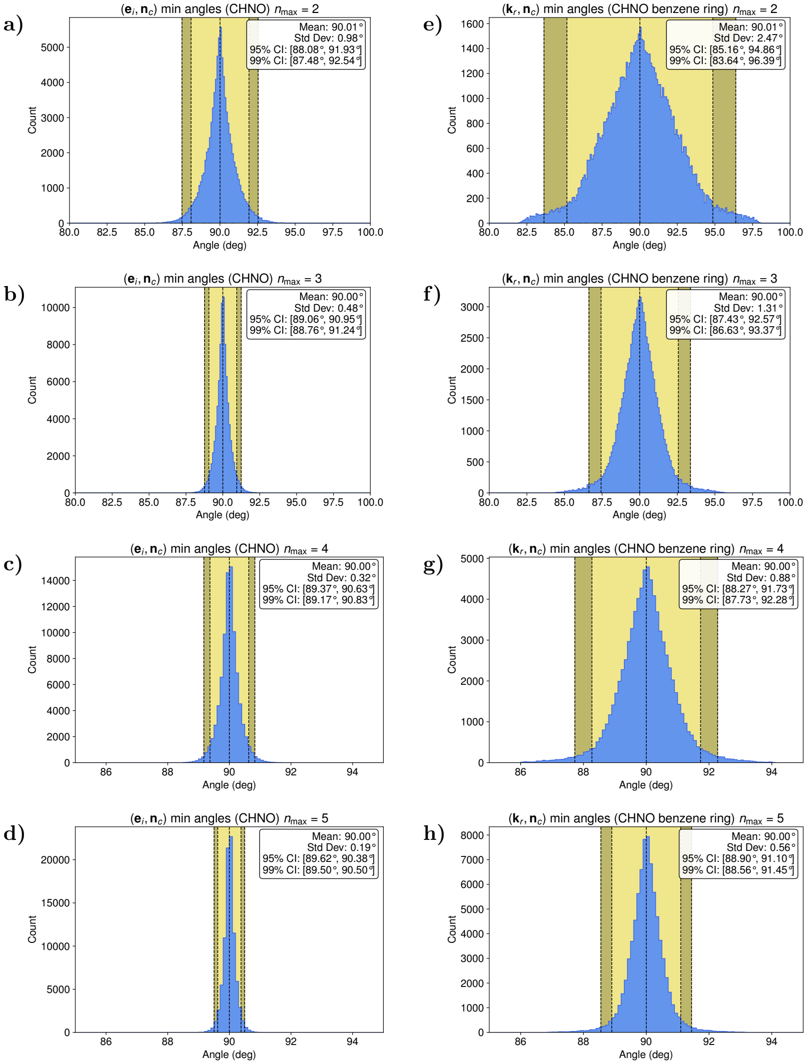


Fig. S.4 a-d) Distributions of the minimum angle formed by the vectors $\mathbf{e}_i$ and $\mathbf{n}_c$ for $n_{\max} = \{2, 3, 4, 5\}$ for all $Z' \leq 5$ structures composed of C, H, N, O atoms. **e-h)** Distributions of the minimum angle formed by the vectors $\mathbf{k}_r$ of the benzene rings and $\mathbf{n}_c$ for $n_{\max} = \{2, 3, 4, 5\}$ for all $Z' \leq 5$ structures composed of C, H, N, O atoms.

3.2 Atomic separations

The analysis of atomic separations for the $Z' \leq 5$ structures in the database revealed that atoms in pairs involving at least one highly electropositive/electronegative atom have the tendency to be found at separations $0.25k_{\mathrm{ZZP}}/|\delta \mathbf{s}_i|$ for at least one direction perpendicular to the vectors $\mathbf{A}$ representing the ZZPs (Fig. S.5).

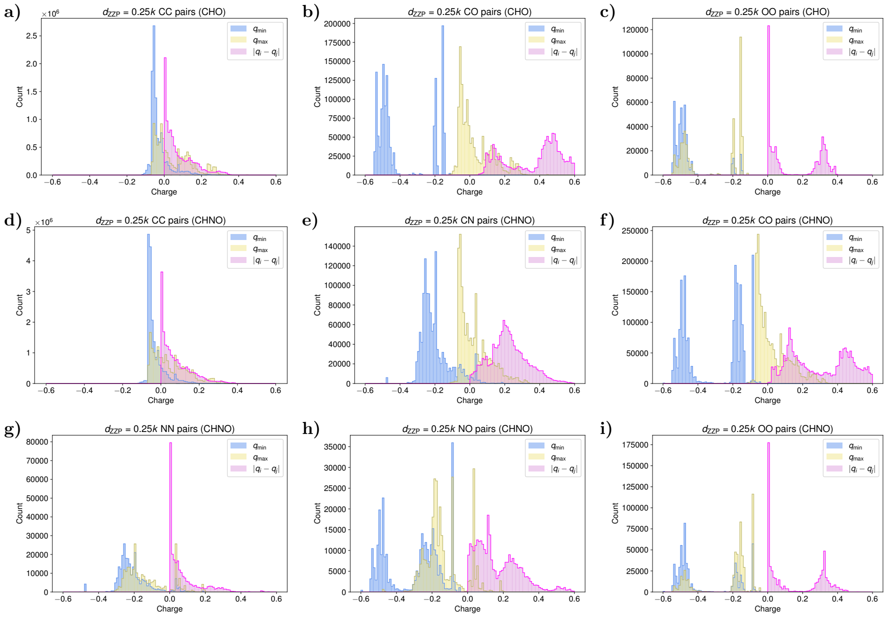


Fig. S.5 Distributions of the minimum charge q_{min}, maximum charge q_{max} and the absolute difference $|q_{\mathrm{max}} - q_{\mathrm{min}}|$ for the possible atomic pairs (excluding hydrogen atoms) in the reference molecule of the $Z' \leq 5$ C, H, O and C, H, N, O structures in the CSD database: **a-c)** The distributions for the CC, CO and OO pairs for the structures composed of C, H, O atoms. **d-i)** The distributions for the CC, CN, CO, NN, NO and OO pairs for the structures composed of C, H, N, O atoms

3.3 Close contacts

The optimal length of the close contacts in a molecular crystal structure depend on the atoms forming the close contact, the nature of the close contact (vdW, hydrogen bond)and also the composition of the molecule. In tables S.1-S.6 we show the 95% confidence intervals for all close contacts in the most common space groups, for structures composed of C, H, O atoms and for structures composed of C, H, N and O atoms.

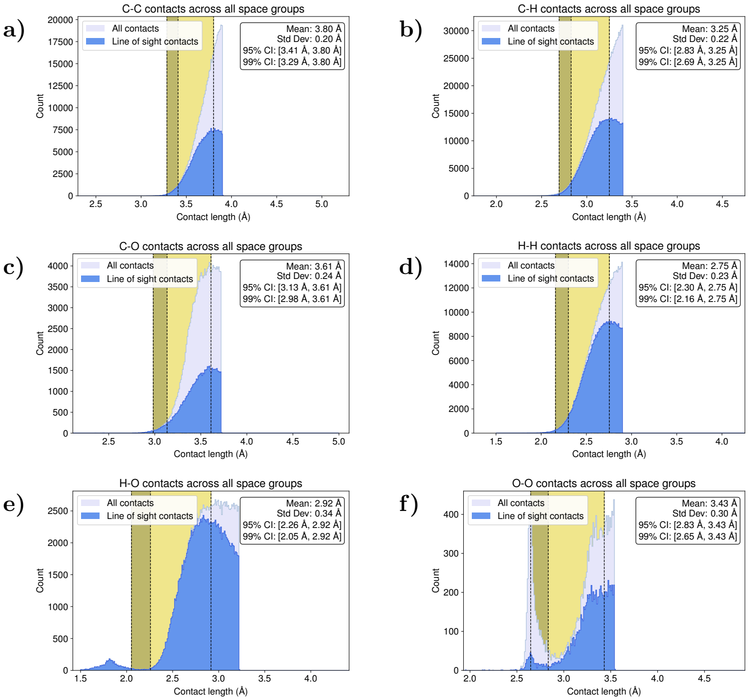


Fig. S.6 Distributions of the line-of-sight contact lengths for the contacts formed by C, H and O atoms for the $Z' \leq 4$ organic molecular crystals with molecular weight ≤ 500 in the CSD database, composed of C, H and O atoms. We can distinguish two patterns, one for the vdW contacts and for the C-C, C-H, C-O, H-H pairs (**a-d**), and a second for the pairs H-O and O-O (**e, f**) that have the ability to form hydrogen bonds.

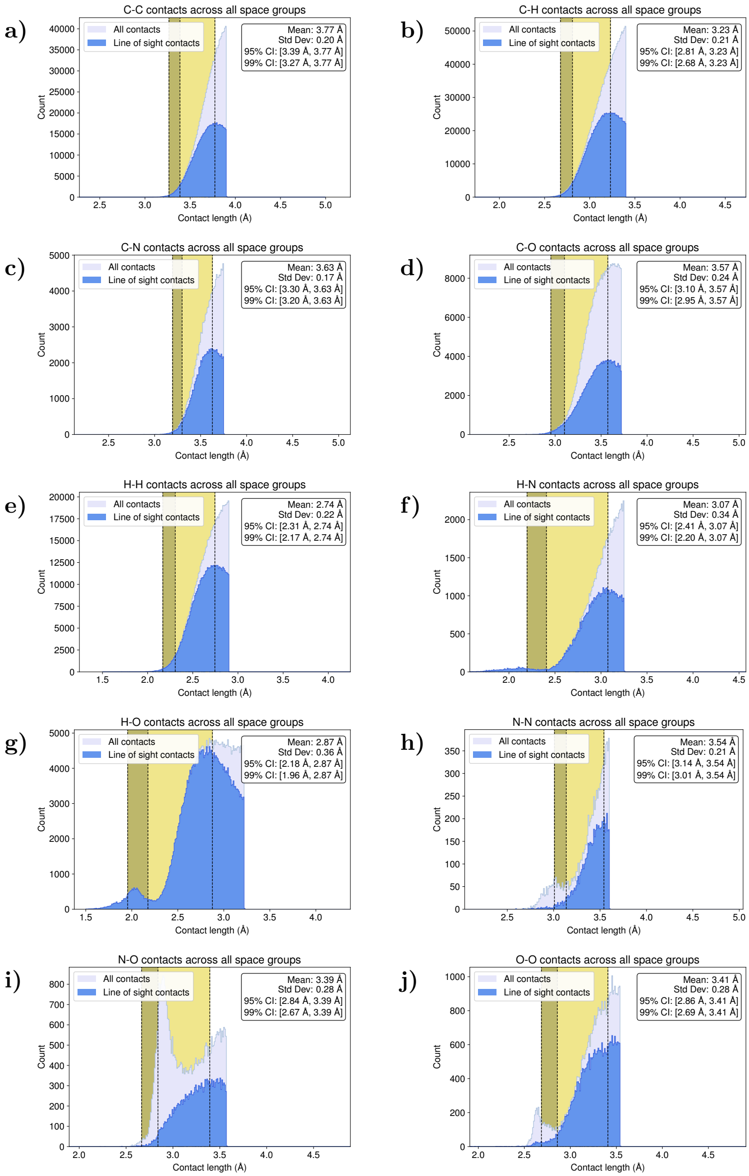


Fig. S.7 Distributions of the line-of-sight contact lengths for the contacts formed by C, H, N and O atoms for the $Z' \leq 5$ organic molecular crystals with molecular weight ≤ 500 in the CSD database, composed of C, H, N and O atoms. We can distinguish two patterns, one for the vdW contacts and for the C-C, C-H, C-N, C-O, H-H pairs (**a-e**), and a second for the pairs H-N, H-O, N-N, N-O and O-O (**f-j**) that have the ability to form hydrogen bonds.

Table S.1 The 95% confidence intervals for the vdW close contacts strength of the structures composed of C, H, and O atoms in the most common space groups.

Space Group	CC	CH	CO
All	(0.00, 0.51)	(0.00, 0.30)	(0.00, 0.27)
C_2	(0.00, 0.16)	(0.00, 0.22)	(0.00, 0.20)
$C2/c$	(0.00, 0.22)	(0.00, 0.26)	(0.00, 0.23)
$C2/m$	–	–	–
Cc	(0.00, 0.73)	(0.00, 0.22)	(0.00, 0.26)
$P-1$	(0.00, 0.54)	(0.00, 0.28)	(0.00, 0.28)
$P2/a$	–	–	–
$P2/c$	(0.00, 0.14)	(0.00, 0.09)	–
$P2/n$	–	(0.00, 0.19)	(0.00, 0.16)
$P2_1$	(0.00, 0.62)	(0.00, 0.36)	(0.00, 0.28)
$P2_1/a$	(0.00, 0.16)	(0.00, 0.24)	(0.00, 0.25)
$P2_1/c$	(0.00, 0.50)	(0.00, 0.27)	(0.00, 0.24)
$P2_1/m$	–	(0.01, 0.10)	–
$P2_1/n$	(0.00, 0.26)	(0.00, 0.25)	(0.00, 0.25)
$P2_12_12$	(0.00, 1.24)	(0.00, 0.31)	(0.00, 0.60)
$P2_12_12_1$	(0.00, 0.42)	(0.00, 0.24)	(0.00, 0.30)
$Pbca$	(0.00, 0.15)	(0.00, 0.20)	(0.00, 0.26)
$Pbcn$	(0.00, 1.83)	(0.00, 1.50)	(0.00, 0.24)
Pc	(0.00, 0.12)	(0.00, 0.20)	(0.00, 0.18)
$Pca2_1$	(0.00, 0.15)	(0.00, 0.19)	(0.00, 0.14)
$Pna2_1$	(0.00, 0.18)	(0.00, 0.19)	(0.00, 0.25)
$Pnma$	–	(0.08, 0.15)	–

Space Group	HH	HO	OO
All	(0.00, 0.29)	(0.00, 0.32)	(0.00, 0.49)
C_2	(0.00, 0.33)	(0.00, 0.30)	(0.05, 0.39)
$C2/c$	(0.00, 0.27)	(0.00, 0.38)	(0.00, 0.51)
$C2/m$	–	–	–
Cc	(0.00, 0.34)	(0.00, 0.33)	(0.05, 0.49)
$P-1$	(0.00, 0.36)	(0.00, 0.33)	(0.00, 0.51)
$P2/a$	–	–	–
$P2/c$	(0.00, 0.12)	(0.05, 0.36)	(0.00, 0.27)
$P2/n$	(0.00, 0.21)	(0.00, 0.38)	(0.00, 0.48)
$P2_1$	(0.00, 0.30)	(0.00, 0.32)	(0.02, 0.41)
$P2_1/a$	(0.00, 0.30)	(0.00, 0.33)	(0.00, 0.25)
$P2_1/c$	(0.00, 0.27)	(0.00, 0.31)	(0.00, 0.42)
$P2_1/m$	–	–	–
$P2_1/n$	(0.00, 0.29)	(0.00, 0.33)	(0.00, 0.44)
$P2_12_12$	(0.00, 0.38)	(0.00, 0.34)	(0.06, 0.53)
$P2_12_12_1$	(0.00, 0.23)	(0.00, 0.33)	(0.01, 0.47)
$Pbca$	(0.00, 0.21)	(0.00, 0.31)	(0.00, 0.47)
$Pbcn$	(0.00, 0.90)	(0.00, 0.32)	(0.01, 0.22)
Pc	(0.00, 0.14)	(0.00, 0.35)	–
$Pca2_1$	(0.00, 0.21)	(0.00, 0.30)	(0.35, 0.36)
$Pna2_1$	(0.00, 0.19)	(0.00, 0.34)	(0.02, 0.26)
$Pnma$	(0.00, 0.10)	(0.00, 0.20)	–

Table S.2 The 95% confidence intervals for the mixed close contacts strength of the structures composed of C, H, and O atoms in the most common space groups. No C-C mixed contacts were identified in the database, and thus they have been excluded from the table

Space Group	CH	CO	HH	HO	OO
All	$(0.00, 0.36)$	$(0.00, 0.23)$	$(0.00, 0.52)$	$(0.00, 0.34)$	$(0.00, 0.48)$
C_2	$(0.00, 0.28)$	$(0.00, 0.17)$	$(0.00, 0.36)$	$(0.00, 0.34)$	$(0.03, 0.43)$
$C2/c$	$(0.00, 0.44)$	$(0.00, 0.18)$	$(0.00, 0.64)$	$(0.00, 0.32)$	$(0.00, 0.49)$
$C2/m$	–	–	–	–	–
Cc	$(0.00, 0.31)$	$(0.00, 0.12)$	$(0.00, 0.23)$	$(0.00, 0.30)$	$(0.02, 0.47)$
$P-1$	$(0.00, 0.40)$	$(0.00, 0.20)$	$(0.00, 0.49)$	$(0.00, 0.28)$	$(0.00, 0.47)$
$P2/a$	–	–	–	–	–
$P2/c$	$(0.00, 0.22)$	–	–	$(0.00, 0.20)$	–
$P2/n$	$(0.00, 0.20)$	–	–	$(0.02, 0.14)$	–
$P2_1$	$(0.00, 0.29)$	$(0.00, 0.21)$	$(0.00, 0.41)$	$(0.00, 0.32)$	$(0.00, 0.47)$
$P2_1/a$	$(0.00, 0.51)$	$(0.00, 0.30)$	$(0.00, 0.24)$	$(0.00, 0.29)$	$(0.00, 0.50)$
$P2_1/c$	$(0.00, 0.40)$	$(0.00, 0.21)$	$(0.00, 0.67)$	$(0.00, 0.35)$	$(0.00, 0.46)$
$P2_1/m$	–	–	–	–	–
$P2_1/n$	$(0.00, 0.38)$	$(0.00, 0.17)$	$(0.00, 0.84)$	$(0.00, 0.44)$	$(0.00, 0.45)$
$P2_12_12$	$(0.00, 0.37)$	$(0.00, 0.35)$	$(0.00, 0.93)$	$(0.00, 0.35)$	$(0.02, 0.48)$
$P2_12_12_1$	$(0.00, 0.30)$	$(0.00, 0.26)$	$(0.00, 0.36)$	$(0.00, 0.35)$	$(0.00, 0.51)$
$Pbca$	$(0.00, 0.44)$	$(0.00, 0.25)$	$(0.00, 0.46)$	$(0.00, 0.29)$	$(0.00, 0.56)$
$Pbcn$	$(0.00, 0.39)$	$(0.00, 0.27)$	$(0.01, 0.06)$	$(0.00, 0.31)$	–
Pc	$(0.01, 0.32)$	–	–	$(0.02, 0.21)$	–
$Pca2_1$	$(0.00, 0.28)$	$(0.03, 0.13)$	$(0.01, 0.19)$	$(0.00, 0.27)$	$(0.14, 0.53)$
$Pna2_1$	$(0.00, 0.36)$	$(0.00, 0.20)$	$(0.00, 0.31)$	$(0.00, 0.29)$	$(0.00, 0.51)$
$Pnma$	$(0.00, 0.38)$	–	$(0.00, 1.66)$	$(0.12, 0.26)$	–

Table S.3 The 95% confidence intervals for the H-bond close contacts strength of the structures composed of C, H, and O atoms in the most common space groups. No C-C, C-H, C-O H-bond contacts were identified in the database, and thus they have been excluded from the table

Space Group	HH	HO	OO
All	$(0.00, 0.54)$	$(0.28, 1.15)$	$(0.06, 0.47)$
C_2	$(0.00, 0.54)$	$(0.28, 1.08)$	$(0.06, 0.44)$
$C2/c$	$(0.00, 0.88)$	$(0.24, 1.23)$	$(0.05, 0.50)$
$C2/m$	$-$	$-$	$-$
Cc	$(0.00, 0.44)$	$(0.42, 1.11)$	$(0.09, 0.47)$
$P-1$	$(0.00, 0.64)$	$(0.27, 1.22)$	$(0.07, 0.50)$
$P2/a$	$-$	$-$	$-$
$P2/c$	$(0.00, 0.16)$	$(0.37, 1.11)$	$(0.00, 0.37)$
$P2/n$	$(0.00, 0.31)$	$(0.43, 1.10)$	$(0.01, 0.48)$
$P2_1$	$(0.00, 0.47)$	$(0.31, 1.08)$	$(0.06, 0.43)$
$P2_1/a$	$(0.00, 0.82)$	$(0.26, 1.29)$	$(0.09, 0.48)$
$P2_1/c$	$(0.00, 0.66)$	$(0.28, 1.21)$	$(0.07, 0.50)$
$P2_1/m$	$-$	$-$	$-$
$P2_1/n$	$(0.00, 0.58)$	$(0.29, 1.17)$	$(0.06, 0.48)$
$P2_12_12$	$(0.00, 0.40)$	$(0.17, 1.19)$	$(0.05, 0.48)$
$P2_12_12_1$	$(0.00, 0.40)$	$(0.27, 1.11)$	$(0.06, 0.44)$
$Pbca$	$(0.00, 0.53)$	$(0.31, 1.24)$	$(0.11, 0.48)$
$Pbcn$	$(0.00, 1.25)$	$(0.42, 1.10)$	$(0.01, 0.46)$
Pc	$-$	$(0.43, 1.10)$	$(0.06, 0.48)$
$Pca2_1$	$(0.00, 0.30)$	$(0.39, 1.09)$	$(0.11, 0.41)$
$Pna2_1$	$(0.00, 0.28)$	$(0.41, 1.06)$	$(0.09, 0.47)$
$Pnma$	$-$	$(0.85, 0.94)$	$(0.14, 0.49)$

Table S.4 The 95% confidence intervals for the vdW close contacts strength of the structures composed of C, H, N and O atoms in the most common space groups.

Space Group	CC	CH	CN	CO	HH
All	$(0.00, 0.47)$	$(0.00, 0.26)$	$(0.00, 0.72)$	$(0.00, 0.28)$	$(0.00, 0.27)$
C_2	$(0.00, 0.16)$	$(0.00, 0.22)$	$(0.00, 0.10)$	$(0.00, 0.27)$	$(0.00, 0.26)$
$C2/c$	$(0.00, 0.86)$	$(0.00, 0.24)$	$(0.00, 0.70)$	$(0.00, 0.34)$	$(0.00, 0.31)$
$C2/m$	–	–	–	–	$(0.04, 0.07)$
Cc	$(0.00, 0.32)$	$(0.00, 0.31)$	$(0.00, 0.10)$	$(0.00, 0.29)$	$(0.00, 0.26)$
$P-1$	$(0.00, 0.47)$	$(0.00, 0.29)$	$(0.00, 0.98)$	$(0.00, 0.28)$	$(0.00, 0.29)$
$P2/a$	–	$(0.00, 0.11)$	–	–	–
$P2/c$	$(0.00, 0.34)$	$(0.00, 0.26)$	$(0.00, 0.10)$	$(0.00, 0.30)$	$(0.00, 0.29)$
$P2/n$	$(0.00, 0.15)$	$(0.00, 0.18)$	–	–	$(0.00, 0.21)$
$P2_1$	$(0.00, 0.36)$	$(0.00, 0.23)$	$(0.00, 0.60)$	$(0.00, 0.32)$	$(0.00, 0.26)$
$P2_1/a$	$(0.00, 0.83)$	$(0.00, 0.40)$	$(0.00, 1.00)$	$(0.00, 0.52)$	$(0.00, 0.34)$
$P2_1/c$	$(0.00, 0.42)$	$(0.00, 0.28)$	$(0.00, 0.75)$	$(0.00, 0.27)$	$(0.00, 0.28)$
$P2_1/m$	–	$(0.00, 0.24)$	–	–	$(0.02, 0.18)$
$P2_1/n$	$(0.00, 0.33)$	$(0.00, 0.24)$	$(0.00, 0.52)$	$(0.00, 0.25)$	$(0.00, 0.25)$
$P2_12_12$	$(0.00, 0.18)$	$(0.00, 0.24)$	–	$(0.00, 0.28)$	$(0.00, 0.51)$
$P2_12_12_1$	$(0.00, 0.53)$	$(0.00, 0.22)$	$(0.00, 0.22)$	$(0.00, 0.25)$	$(0.00, 0.23)$
$Pbca$	$(0.00, 0.20)$	$(0.00, 0.21)$	$(0.00, 0.20)$	$(0.00, 0.24)$	$(0.00, 0.21)$
$Pbcn$	$(0.00, 0.21)$	$(0.00, 0.24)$	$(0.00, 0.22)$	$(0.00, 0.20)$	$(0.00, 0.28)$
Pc	$(0.00, 0.18)$	$(0.00, 0.23)$	$(0.00, 0.19)$	$(0.00, 0.22)$	$(0.00, 0.30)$
$Pca2_1$	$(0.00, 0.15)$	$(0.00, 0.21)$	$(0.00, 0.17)$	$(0.00, 0.26)$	$(0.00, 0.24)$
$Pna2_1$	$(0.00, 0.38)$	$(0.00, 0.21)$	$(0.00, 0.15)$	$(0.00, 0.25)$	$(0.00, 0.23)$
$Pnma$	$(0.00, 1.09)$	$(0.00, 1.02)$	$(0.00, 0.13)$	$(0.00, 0.15)$	$(0.00, 1.58)$

Space Group	HN	HO	NN	NO	OO
All	$(0.00, 0.31)$	$(0.00, 0.35)$	$(0.00, 0.80)$	$(0.00, 0.33)$	$(0.00, 0.41)$
C_2	$(0.00, 0.23)$	$(0.00, 0.35)$	–	$(0.00, 0.35)$	$(0.00, 0.38)$
$C2/c$	$(0.00, 0.30)$	$(0.00, 0.36)$	$(0.00, 0.29)$	$(0.00, 0.30)$	$(0.00, 0.41)$
$C2/m$	$(0.00, 0.05)$	$(0.12, 0.33)$	–	–	–
Cc	$(0.00, 0.30)$	$(0.00, 0.35)$	$(0.00, 0.08)$	$(0.00, 0.35)$	$(0.00, 0.21)$
$P-1$	$(0.00, 0.34)$	$(0.00, 0.35)$	$(0.00, 0.69)$	$(0.00, 0.34)$	$(0.00, 0.47)$
$P2/a$	–	$(0.00, 0.42)$	–	–	–
$P2/c$	$(0.00, 0.29)$	$(0.00, 0.39)$	–	$(0.00, 0.88)$	$(0.00, 0.44)$
$P2/n$	–	$(0.00, 0.35)$	–	$(0.00, 0.13)$	$(0.12, 0.38)$
$P2_1$	$(0.00, 0.30)$	$(0.00, 0.35)$	$(0.00, 1.56)$	$(0.00, 0.30)$	$(0.00, 0.47)$
$P2_1/a$	$(0.00, 0.29)$	$(0.00, 0.40)$	$(0.02, 0.13)$	$(0.00, 0.99)$	$(0.00, 0.57)$
$P2_1/c$	$(0.00, 0.33)$	$(0.00, 0.35)$	$(0.00, 1.02)$	$(0.00, 0.33)$	$(0.00, 0.39)$
$P2_1/m$	$(0.02, 0.16)$	$(0.01, 0.22)$	–	–	–
$P2_1/n$	$(0.00, 0.29)$	$(0.00, 0.35)$	$(0.00, 0.71)$	$(0.00, 0.32)$	$(0.00, 0.37)$
$P2_12_12$	$(0.00, 0.35)$	$(0.00, 0.46)$	–	$(0.01, 0.03)$	$(0.00, 0.52)$
$P2_12_12_1$	$(0.00, 0.29)$	$(0.00, 0.35)$	$(0.00, 0.26)$	$(0.00, 0.32)$	$(0.00, 0.47)$
$Pbca$	$(0.00, 0.27)$	$(0.00, 0.36)$	$(0.00, 0.32)$	$(0.00, 0.29)$	$(0.00, 0.28)$
$Pbcn$	$(0.00, 0.30)$	$(0.00, 0.33)$	–	$(0.00, 0.42)$	$(0.00, 0.52)$
Pc	$(0.00, 0.36)$	$(0.00, 0.37)$	–	$(0.00, 0.30)$	$(0.00, 0.42)$
$Pca2_1$	$(0.00, 0.33)$	$(0.00, 0.36)$	$(0.00, 0.22)$	$(0.00, 0.32)$	$(0.00, 0.38)$
$Pna2_1$	$(0.00, 0.32)$	$(0.00, 0.36)$	$(0.00, 0.27)$	$(0.00, 0.32)$	$(0.00, 0.36)$
$Pnma$	$(0.00, 0.43)$	$(0.00, 0.63)$	–	$(0.00, 0.23)$	–

Table S.5 The 95% confidence intervals for the Mixed close contacts strength of the structures composed of C, H, N and O atoms in the most common space groups.

Space Group	CC	CH	CN	CO	HH
All	–	$(0.00, 0.40)$	$(0.00, 0.67)$	$(0.00, 0.30)$	$(0.00, 0.43)$
C_2	–	$(0.00, 0.34)$	$(0.07, 0.08)$	$(0.00, 0.23)$	$(0.00, 0.42)$
$C2/c$	–	$(0.00, 0.41)$	$(0.00, 0.17)$	$(0.00, 0.43)$	$(0.00, 0.53)$
$C2/m$	–	–	–	–	–
Cc	–	$(0.00, 0.36)$	$(0.00, 0.11)$	$(0.00, 0.19)$	$(0.00, 0.22)$
$P-1$	–	$(0.00, 0.40)$	$(0.00, 0.87)$	$(0.00, 0.27)$	$(0.00, 0.54)$
$P2/a$	–	$(0.17, 0.24)$	–	–	–
$P2/c$	–	$(0.00, 0.32)$	$(0.00, 0.03)$	$(0.00, 0.12)$	$(0.00, 0.50)$
$P2/n$	–	$(0.00, 0.28)$	–	$(0.00, 0.15)$	$(0.02, 0.12)$
$P2_1$	–	$(0.00, 0.36)$	$(0.00, 0.10)$	$(0.00, 0.30)$	$(0.00, 0.44)$
$P2_1/a$	–	$(0.00, 0.52)$	$(0.00, 0.37)$	$(0.00, 0.23)$	$(0.00, 0.48)$
$P2_1/c$	–	$(0.00, 0.41)$	$(0.00, 0.81)$	$(0.00, 0.33)$	$(0.00, 0.44)$
$P2_1/m$	–	–	–	–	–
$P2_1/n$	–	$(0.00, 0.40)$	$(0.00, 0.13)$	$(0.00, 0.32)$	$(0.00, 0.41)$
$P2_12_12$	–	$(0.00, 0.37)$	–	$(0.00, 0.15)$	$(0.00, 0.46)$
$P2_12_12_1$	–	$(0.00, 0.37)$	$(0.00, 0.14)$	$(0.00, 0.21)$	$(0.00, 0.31)$
$Pbca$	–	$(0.00, 0.38)$	$(0.00, 0.10)$	$(0.00, 0.38)$	$(0.00, 0.29)$
$Pbcn$	–	$(0.00, 0.37)$	$(0.00, 0.10)$	$(0.00, 0.22)$	$(0.00, 0.24)$
Pc	–	$(0.00, 0.43)$	$(0.00, 0.12)$	$(0.00, 0.17)$	$(0.00, 0.28)$
$Pca2_1$	–	$(0.00, 0.42)$	$(0.00, 0.09)$	$(0.00, 0.24)$	$(0.00, 0.22)$
$Pna2_1$	–	$(0.00, 0.37)$	$(0.00, 0.13)$	$(0.00, 0.16)$	$(0.00, 0.32)$
$Pnma$	–	$(0.00, 0.59)$	–	$(0.00, 0.24)$	–

Space Group	HN	HO	NN	NO	OO
All	$(0.00, 0.43)$	$(0.00, 0.36)$	$(0.00, 1.18)$	$(0.00, 0.34)$	$(0.00, 0.49)$
C_2	$(0.00, 0.16)$	$(0.00, 0.35)$	–	$(0.03, 0.34)$	$(0.03, 0.44)$
$C2/c$	$(0.00, 0.25)$	$(0.00, 0.34)$	$(0.00, 0.24)$	$(0.00, 0.37)$	$(0.00, 0.60)$
$C2/m$	–	–	–	–	–
Cc	$(0.00, 0.25)$	$(0.00, 0.36)$	$(0.04, 0.09)$	$(0.00, 0.29)$	$(0.00, 0.27)$
$P-1$	$(0.00, 0.46)$	$(0.00, 0.36)$	$(0.00, 1.40)$	$(0.00, 0.36)$	$(0.00, 0.52)$
$P2/a$	–	–	–	–	–
$P2/c$	$(0.00, 0.24)$	$(0.00, 0.41)$	–	$(0.00, 0.10)$	–
$P2/n$	$(0.00, 0.03)$	$(0.00, 0.34)$	–	–	$(0.13, 0.52)$
$P2_1$	$(0.00, 0.42)$	$(0.00, 0.37)$	$(0.00, 0.13)$	$(0.00, 0.33)$	$(0.00, 0.49)$
$P2_1/a$	$(0.03, 0.20)$	$(0.00, 0.43)$	–	$(0.00, 0.37)$	$(0.03, 0.51)$
$P2_1/c$	$(0.00, 0.58)$	$(0.00, 0.35)$	$(0.00, 1.57)$	$(0.00, 0.31)$	$(0.00, 0.45)$
$P2_1/m$	–	–	–	–	–
$P2_1/n$	$(0.00, 0.26)$	$(0.00, 0.37)$	$(0.00, 0.29)$	$(0.00, 0.32)$	$(0.00, 0.45)$
$P2_12_12$	$(0.00, 0.19)$	$(0.00, 0.36)$	–	$(0.04, 0.35)$	$(0.00, 0.48)$
$P2_12_12_1$	$(0.00, 0.24)$	$(0.00, 0.34)$	$(0.05, 0.31)$	$(0.00, 0.33)$	$(0.00, 0.54)$
$Pbca$	$(0.00, 0.26)$	$(0.00, 0.35)$	$(0.00, 0.19)$	$(0.00, 0.36)$	$(0.00, 0.50)$
$Pbcn$	$(0.00, 0.24)$	$(0.00, 0.35)$	–	$(0.00, 0.55)$	$(0.00, 0.50)$
Pc	$(0.00, 0.23)$	$(0.00, 0.31)$	–	$(0.00, 0.27)$	$(0.00, 0.48)$
$Pca2_1$	$(0.00, 0.28)$	$(0.00, 0.33)$	$(0.03, 0.11)$	$(0.00, 0.31)$	$(0.00, 0.46)$
$Pna2_1$	$(0.00, 0.26)$	$(0.00, 0.37)$	$(0.06, 0.17)$	$(0.00, 0.29)$	$(0.00, 0.40)$
$Pnma$	–	$(0.00, 0.21)$	–	$(0.00, 0.18)$	–

Table S.6 The 95% confidence intervals for the H-bond close contacts strength of the structures composed of C, H, N and O atoms in the most common space groups.

Space Group	CC	CH	CN	CO	HH
All	–	–	–	–	$(0.00, 0.52)$
C_2	–	–	–	–	$(0.00, 0.72)$
$C2/c$	–	–	–	–	$(0.00, 0.68)$
$C2/m$	–	–	–	–	–
Cc	–	–	–	–	$(0.00, 0.26)$
$P-1$	–	–	–	–	$(0.00, 0.48)$
$P2/a$	–	–	–	–	–
$P2/c$	–	–	–	–	$(0.00, 0.76)$
$P2/n$	–	–	–	–	$(0.00, 0.77)$
$P2_1$	–	–	–	–	$(0.00, 0.45)$
$P2_1/a$	–	–	–	–	$(0.00, 0.43)$
$P2_1/c$	–	–	–	–	$(0.00, 0.64)$
$P2_1/m$	–	–	–	–	–
$P2_1/n$	–	–	–	–	$(0.00, 0.43)$
$P2_12_12$	–	–	–	–	$(0.00, 0.56)$
$P2_12_12_1$	–	–	–	–	$(0.00, 0.37)$
$Pbca$	–	–	–	–	$(0.00, 0.56)$
$Pbcn$	–	–	–	–	$(0.00, 0.88)$
Pc	–	–	–	–	$(0.00, 0.57)$
$Pca2_1$	–	–	–	–	$(0.00, 0.29)$
$Pna2_1$	–	–	–	–	$(0.00, 0.26)$
$Pnma$	–	–	–	–	–

Space Group	HN	HO	NN	NO	OO
All	$(0.12, 1.19)$	$(0.16, 1.13)$	$(0.00, 0.30)$	$(0.00, 0.46)$	$(0.05, 0.52)$
C_2	$(0.13, 1.14)$	$(0.22, 1.08)$	$(0.02, 0.17)$	$(0.00, 0.42)$	$(0.07, 0.46)$
$C2/c$	$(0.12, 1.18)$	$(0.14, 1.13)$	$(0.00, 0.28)$	$(0.00, 0.46)$	$(0.04, 0.53)$
$C2/m$	–	–	–	–	–
Cc	$(0.14, 1.20)$	$(0.11, 1.10)$	$(0.00, 0.40)$	$(0.00, 0.47)$	$(0.05, 0.50)$
$P-1$	$(0.16, 1.21)$	$(0.15, 1.14)$	$(0.00, 0.30)$	$(0.00, 0.49)$	$(0.04, 0.54)$
$P2/a$	$(0.52, 0.57)$	–	–	–	–
$P2/c$	$(0.02, 1.14)$	$(0.07, 1.13)$	$(0.02, 0.16)$	$(0.00, 0.51)$	$(0.03, 0.54)$
$P2/n$	$(0.09, 1.11)$	$(0.13, 1.19)$	$(0.00, 0.31)$	$(0.02, 0.34)$	$(0.11, 0.51)$
$P2_1$	$(0.12, 1.14)$	$(0.22, 1.08)$	$(0.00, 0.36)$	$(0.00, 0.41)$	$(0.07, 0.47)$
$P2_1/a$	$(0.25, 1.30)$	$(0.21, 1.15)$	$(0.00, 0.27)$	$(0.01, 0.47)$	$(0.00, 0.52)$
$P2_1/c$	$(0.10, 1.17)$	$(0.16, 1.15)$	$(0.00, 0.29)$	$(0.00, 0.46)$	$(0.05, 0.54)$
$P2_1/m$	–	–	–	–	–
$P2_1/n$	$(0.12, 1.19)$	$(0.16, 1.13)$	$(0.00, 0.30)$	$(0.00, 0.47)$	$(0.05, 0.52)$
$P2_12_12$	$(0.02, 1.15)$	$(0.24, 1.09)$	$(0.00, 0.29)$	$(0.03, 0.35)$	$(0.10, 0.47)$
$P2_12_12_1$	$(0.15, 1.13)$	$(0.20, 1.12)$	$(0.00, 0.29)$	$(0.00, 0.41)$	$(0.06, 0.49)$
$Pbca$	$(0.07, 1.22)$	$(0.15, 1.12)$	$(0.00, 0.36)$	$(0.00, 0.46)$	$(0.05, 0.54)$
$Pbcn$	$(0.19, 1.20)$	$(0.18, 1.14)$	$(0.00, 0.29)$	$(0.00, 0.48)$	$(0.06, 0.49)$
Pc	$(0.00, 1.41)$	$(0.06, 1.10)$	$(0.04, 0.19)$	$(0.00, 0.46)$	$(0.05, 0.56)$
$Pca2_1$	$(0.08, 1.21)$	$(0.11, 1.11)$	$(0.00, 0.33)$	$(0.00, 0.46)$	$(0.05, 0.53)$
$Pna2_1$	$(0.13, 1.20)$	$(0.12, 1.15)$	$(0.00, 0.33)$	$(0.00, 0.48)$	$(0.04, 0.56)$
$Pnma$	$(0.56, 0.75)$	$(0.00, 1.35)$	$(0.15, 0.19)$	$(0.00, 0.43)$	$(0.22, 0.62)$

4 Search for the aspirin polymorphs - example structures

During the flexible search of aspirin polymorphs, several structures are discarded at each step of the filtering process, while the structures making it to the final pool undergo 2 optimization steps during the process, slightly altering the structure of the unit cells. CIF files for low vdW free volume structures rejected for this (and all other) systems studied in the main manuscript have been provided as additional supporting material.

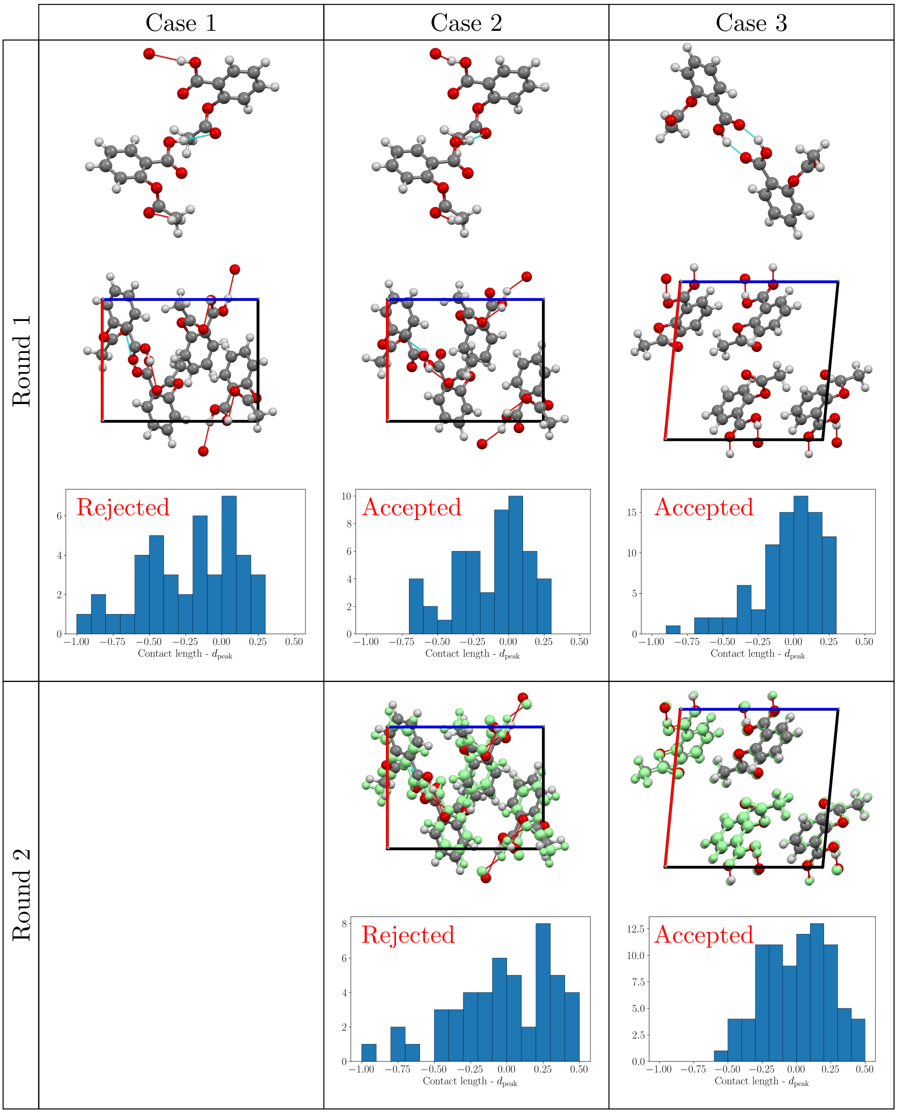
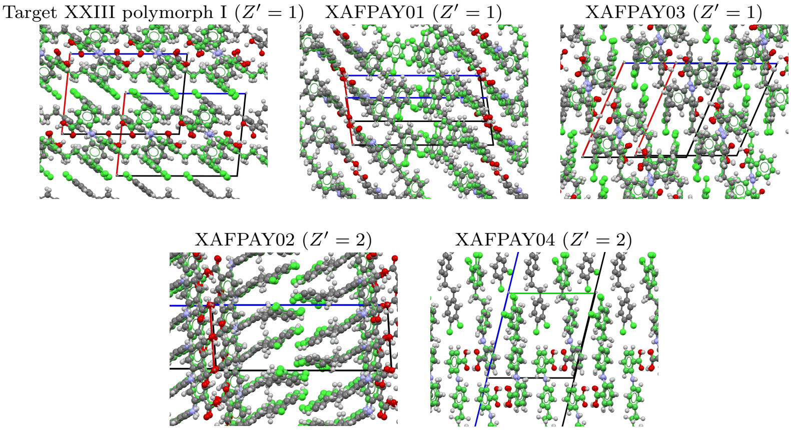


Fig. S.8 The figure illustrates three different examples of contact filtering in the case of aspirin structures. In Round 1, the structures are optimized to achieve adherence of the heavy atoms to the ZZPs. In Round 2, the structures are optimized for close contacts. Cases 1 and 2 involve identical conformations within the same unit cell geometry but with different placements of the reference molecule. In both cases, a hydrogen bond is formed between the carboxyl fragment and the terminal oxygen of the ester fragment. In Case 1, the distribution of close contacts, measured by the deviation of the length from the peak of the distribution for each atomic pair, includes several below the accepted limit, resulting in the structure being discarded in Round 1. Conversely, in Case 2, the contact distribution is accepted in Round 1. However, after the optimization step for the close contacts, several very short contacts appear, leading to the structure being discarded. Case 3 involves a different conformation and cell geometry, forming hydrogen bonds between the oxygen atoms in the carboxyl group. In both filtering rounds, the contact distribution is consistent with that in the CSD database, allowing the structure to be accepted into the final pool.

5 Search for the target XXIII polymorphs

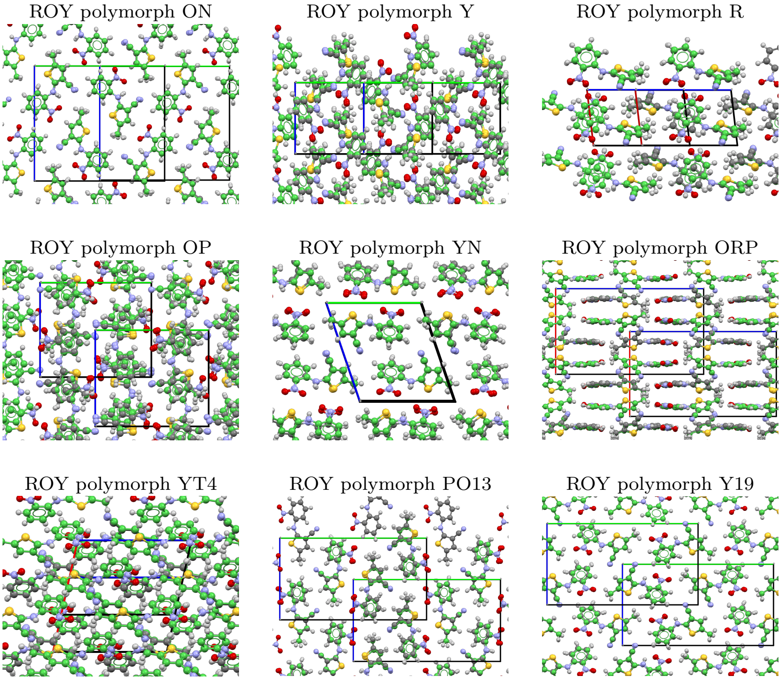


Target XXIII polymorph I ($Z' = 1$) XAFPAY01 ($Z' = 1$) XAFPAY03 ($Z' = 1$)

XAFPAY02 ($Z' = 2$) XAFPAY04 ($Z' = 2$)

Fig. S.9 Overlays of the 3 $Z' = 1$ and 2 $Z' = 2$ target XXIII polymorphs identified in the search.

6 Search for the ROY polymorphs



ROY polymorph ON

ROY polymorph Y

ROY polymorph R

ROY polymorph OP

ROY polymorph YN

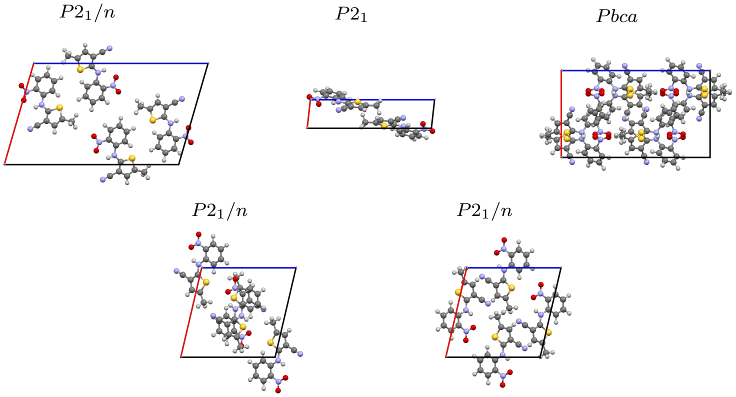
ROY polymorph ORP

ROY polymorph YT4

ROY polymorph PO13

ROY polymorph Y19

Fig. S.10 Overlays of the 9 $Z' = 1$ ROY polymorphs identified in the search.



$P2_1/n$

$P2_1$

$Pbca$

$P2_1/n$

$P2_1/n$

Fig. S.11 The five unique ROY structures that appeared in the final pool of the search.

7 Cost function vs vdWFV landscapes

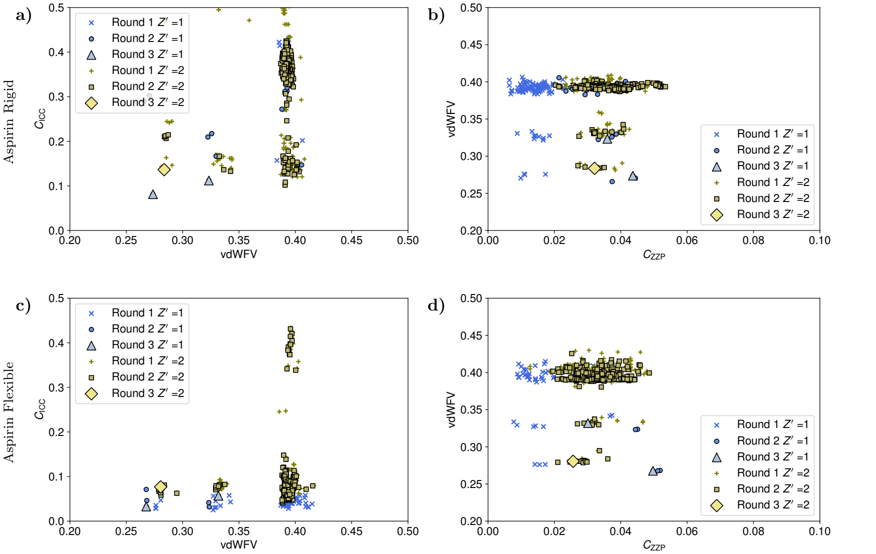


Fig. S.12 a, b) Landscapes of the vdWFV against the cost functions C_{ICC} and C_{ZZP} for the rigid aspirin search. **c, d)** Landscapes of the vdWFV against the cost functions C_{ICC} and C_{ZZP} for the flexible aspirin search.

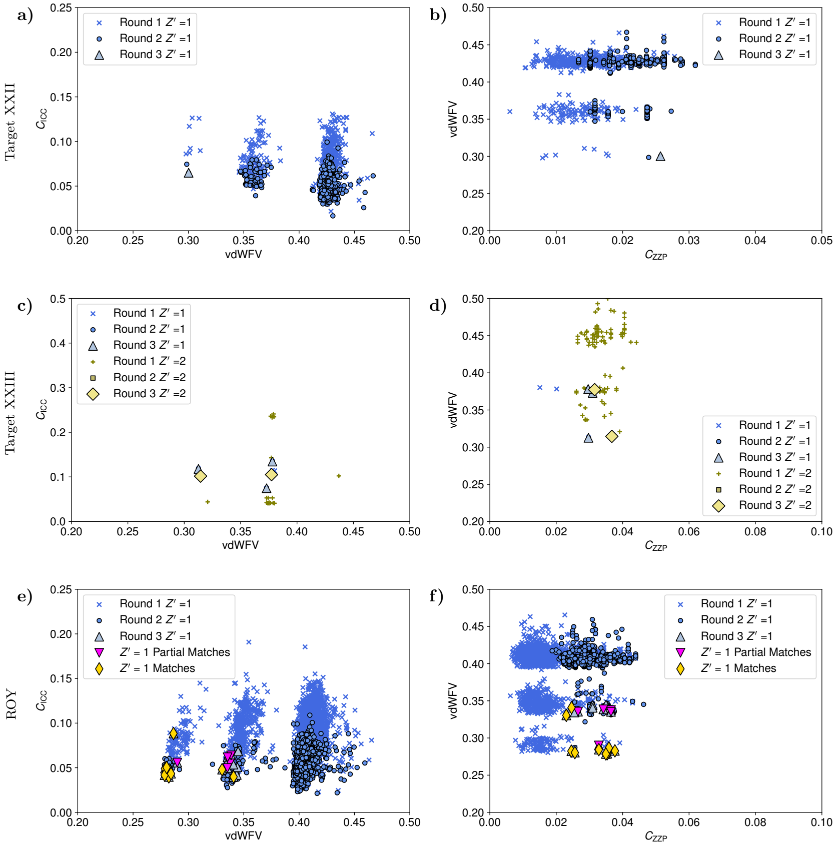


Fig. S.13 **a, b)** Landscapes of the vdWFV against the cost functions C_{ICC} and C_{ZZP} for the target XXII search. **c, d)** Landscapes of the vdWFV against the cost functions C_{ICC} and C_{ZZP} for the target XXIII search. **e, f)** Landscapes of the vdWFV against the cost functions C_{ICC} and C_{ZZP} for the ROY search.