ENTROPIC TRUST REGION FOR DENSEST CRYSTALLOGRAPHIC SYMMETRY GROUP PACKINGS [*]

MILOSLAV TORDA [†], JOHN Y. GOULERMAS [‡], ROLAND PÚČEK [§], AND VITALIY KURLIN [‡]

Abstract. Molecular crystal structure prediction (CSP) seeks the most stable periodic structure given the chemical composition of a molecule and pressure-temperature conditions. Modern CSP solvers use global optimization methods to search for structures with minimal free energy within a complex energy landscape induced by intermolecular potentials. A major caveat of these methods is that initial configurations are random, making thus the search susceptible to convergence at local minima. Providing initial configurations that are densely packed with respect to the geometric representation of a molecule can significantly accelerate CSP. Motivated by these observations, we define a class of periodic packings restricted to crystallographic symmetry groups (CSG) and design a search method for the densest CSG packings in an information-geometric framework. Since the CSG induce a toroidal topology on the configuration space, a non-euclidean trust region method is performed on a statistical manifold consisting of probability distributions defined on an n-dimensional flat unit torus by extending the multivariate von Mises distribution. Introducing an adaptive quantile reformulation of the fitness function into the optimization schedule provides the algorithm with a geometric characterization through local dual geodesic flows. Moreover, we examine the geometry of the adaptive selection-quantile defined trust region and show that the algorithm performs a maximization of stochastic dependence among elements of the extended multivariate von Mises distributed random vector. We experimentally evaluate its behaviour and performance on various densest packings of convex polygons in 2-dimensional CSGs for which optimal solutions are known. Additionally, we demonstrate the application of the densest CSG packings in the pentacene thin-film CSP.

Key words. Crystal structure prediction, Directional statistics, Geometric packing, Information - geometric optimization, Evolutionary strategies.

AMS subject classifications. 68W50, 90C56, 65C05, 52C15, 62H11, 74E15, 90C90, 94A17, 53B12

1. Introduction. The work presented here is motivated by the problem of Crystal Structure Prediction (CSP), in which, given some molecular shape K, the goal is to predict a synthesizable periodic structure. Such a periodic structure may consist of several copies of K within a unit cell formation (parallelepiped) that is periodically repeated along the three directions. Figure 1 exemplifies such a formation with a 2D pentagonal crystal. CSP traditionally starts from an almost random configuration of molecules in a random unit cell and attempts to optimize a complex energy function depending on the given molecular structure and numerous problem parameters such as pressure-temperature conditions.

Current CSP approaches have two main computational caveats. The first is energy computation, where either one of many empirical potentials needs to be chosen or computationally expensive, but precise density-functional theory calculations are used. The second is that the energy functions induce complicated energy landscapes

[†] Leverhulme Research Centre for Functional Materials Design, Department of Computer Science, University of Liverpool, Liverpool, UK (miloslav.torda@liverpool.ac.uk)

[‡] Department of Computer Science, University of Liverpool, Liverpool, UK (j.y.goulermas@liverpool.ac.uk, vitaliy.kurlin@liverpool.ac.uk)

[§] Department of Mathematics, Friedrich Schiller University Jena, Jena, DE (roland.pucek@uni-jena.de)
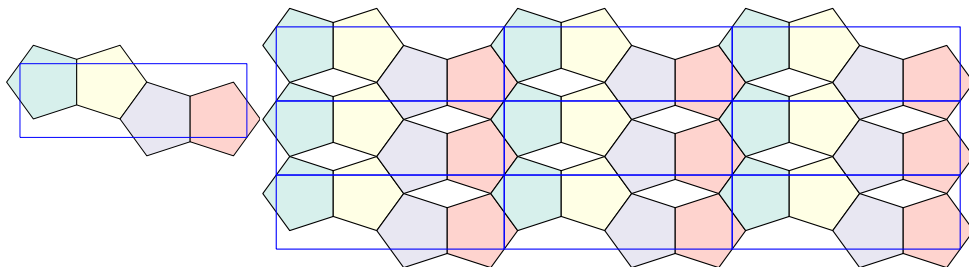
Figure 1: The 2D periodic structure with the $p2mg$ plane group symmetry (i.e., consisting of a 2-fold rotational symmetry operation, glide reflections and mirror reflections along two mirror planes. Colors represent these symmetry operations modulo lattice translations.) where K is a regular convex pentagon with the packing density of approximately 0.8541019. (Left) A single unit cell. (Right) 15 unit cells.

[79], increasing the likelihood of the global search methods converging to a local minimum basin and leading to over-prediction [59]. The standard output of CSP computations is thousands of theoretical polymorphic structures, each representing some local optimum of the energy landscape [87]. Afterwards, data analytic tools are employed to identify metastable structures.

In this work, we propose an approach based on discrete geometry to facilitate the CSP workflow by providing reasonable initial periodic configurations. Specifically, a polytope representation is assigned to a molecule based on its intrinsic properties [66, 80], and the highest density packings of hypothetical structures are generated. These hypothetical configurations are then used as starting positions for the usual CSP energy minimizations, thus reducing the computational burden only to local explorations of the energy landscape.

Geometric packings are well-studied objects in discrete and computational geometry [76] and are fundamental in solid-state physics modelling [75]. Although molecular crystals are considered to be embeddings in the 3D euclidean space, following the currently high interest in 2D materials [23], the same approach can also be employed here. That is, given a representation of a molecule by a polygon, the aim is to acquire the configuration that maximizes packing density and subsequently use this as a starting configuration in classical CSP workflows. Moreover, not only generally densest packings but also lower density but higher symmetry crystal structures that maximize packing density among a particular isomorphism class of periodic structures can point to possible, stable crystal phases. For example, Figure 1 demonstrates a crystallization pattern of pentagonal proteins on lipid mono-layers [81] by densest packing of pentagon when the configuration space is restricted to the $p2mg$ plane group.

This approach is well justified experimentally [22, 26] and by previous molecular dynamics simulations using force-field methods [55, 88]. Furthermore, the crystallization conjecture states that in the Euclidean space of dimensions two and three, the ground state energy of systems of interacting particles forms periodic configurations in the thermodynamic limit [46]. [74] proved the equivalence between the crystallization conjecture for mono-atomic systems in two dimensions and the densest disc packing for a class of Lennard-Jones-like energy potentials. Later, [28] proved the face-centred-cubic sphere packing model's optimality in terms of energy minimization using an additional three-body potential. Even though there are no such results for molecular systems, the usual correlation between packing density maximization and energy minimization suggests an equivalence between the densest packings of

polytopes and the crystallization conjecture, at least for some molecular crystals.

Despite having attracted the interest of various scientific communities for centuries now, constructing the densest packings of a set of given geometric entities is a notoriously hard problem, and only a few optimal solutions are known. In the 3D euclidean space, these include the general packing of the sphere [37] and the truncated rhombic dodecahedron [13] and the existence of an algorithm to construct the densest lattice packings of convex polygons [10]. In the 2D euclidean plane, known general packings include that of the disk [77] and the pentagon [36], algorithms to construct the densest packing of centrally symmetric convex polygons [77], and algorithms for the densest lattice packings and double lattice packings of convex polygons [54, 53]. However, all aforementioned construction methods are tailored to a specific geometric shape. We aim to construct the densest packings for a large class of objects and symmetry groups using a robust and generic method without the restriction to given shapes. To this extent, we have developed an optimization framework based on the natural gradient method [1] used in evolution strategies [38, 85] which are instances of a general information-geometric optimization framework [58].

Although our proposed optimization system was specifically developed to search for the densest crystallographic packings, it can also be applied in classical CSP computations by interfacing with force field methods [29] or density functional theory calculations [35]. This can be achieved by replacing the maximizing packing density objective with minimizing free energy. In essence, due to the generic design of the entropic trust region based optimization, it can be used to solve any bounded and constrained black-box optimization problem.

Since the currently established molecular crystal model is that of a crystallographic symmetry group (CSG), we restrict our search to the 230 space groups and 17 plane groups [15] and define a new subproblem of the general packing problem [64], the CSG packing. The periodic boundary conditions inherent to CSGs, innately induce a toroidal topology on the packing configuration space. We exploit this property by performing natural gradient ascent on a statistical manifold composed of probability distributions on an nD torus using an extension of the multivariate von Mises model for directional data [50]. This effectively removes the optimization problem boundaries, which pose a considerable difficulty for many optimization methods when the solution lies on the configuration space boundaries. Moreover, by starting from the uniform distribution on the torus, we remove the algorithm's dependence on the initial configuration.

The manuscript is divided into four parts. First, Section 2 introduces CSG packings and the related densest CSG packing problem. Section 3 introduces the entropic trust region method and the extended multivariate von Mises distribution and presents the exponential family reformulation for the extended multivariate von Mises distribution. We also introduce the adaptive selection quantile to facilitate the natural gradient ascent. Furthermore, using its connection with the proximal entropic method [73], we examine the geometry of the entropic trust region based exponential family adaptive selection quantile and establish that the algorithm defines, in fact, parallel Riemannian gradient flows between two statistical submanifolds of an ambient manifold; namely, one maximizing entropy [68] and another maximizing multi-information [72]. Finally, the section introduces a method for refining solutions based on the localization of the search in a subspace of the initial configuration space. Section 4 contains various experiments that investigate the behaviour and performance of the entropic trust region method on the densest $p2$ packing of a regular octagon. Results demonstrate higher accuracy than in previous Monte–Carlo methods [4, 67]. Lastly,

in Section 5, we demonstrate the application of our algorithm on the densest packings of a polygonal representation of pentacene in a few plane groups to determine crystal structures of pentacene thin films found in the literature. Many additional experiments and mathematical, algorithmic and implementation details are included in the supplementary material for this manuscript.

2. Densest CSG packings. Even though regarding CSP applications, we are interested in dimensions 2 and 3, we introduce CSGs for general n dimensions following [52].

Let E_n denote the group of all isometries in $\mathbb{R}^n$ that is all length preserving transformations of $\mathbb{R}^n$, and O_n and T_n denote orthogonal and translational subgroups of E_n, respectively. Specifically $O_n = \{\mathbf{R} \mid \mathbf{R}^\mathsf{T}\mathbf{R} = \mathbf{R}\mathbf{R}^\mathsf{T} = \mathbf{I}; \mathbf{R} \in \mathbb{R}^{n \times n}\}$ is the set of all rotations and rotation-inversions that preserve one fixed point $\mathbf{p} \in \mathbb{R}^n$, and $T_n = \{T_\mathbf{a} \mid \mathbf{p}T_\mathbf{a} = \mathbf{p} + \mathbf{a} ; \mathbf{a} \in \mathbb{R}^n\}$ is the set of all translations.

The nD lattice group

$$(2.1) \qquad L = \{T_\mathbf{a} \mid \mathbf{a} = u_1\mathbf{b}_1 + \ldots + u_n\mathbf{b}_n ; \ u_1, \ldots, u_n \in \mathbb{Z}\},$$

is a discrete subgroup of T_n where $\mathbf{b}_1, \ldots, \mathbf{b}_n \in \mathbb{R}^n$ are the basic vectors. We denote the orbit of a point $\mathbf{p}$ under the action of L by $\Lambda = \{\mathbf{p}T_\mathbf{a} \mid T_\mathbf{a} \in L\}$ (the geometric lattice) and the convex hull of basic vectors by $\bar{\Lambda} = \mathrm{conv}\{c_1\mathbf{b}_1 + \ldots + c_n\mathbf{b}_n ; c_1, \ldots, c_n \in [0; 1]\}$ (the primitive cell). The crystallographic point group is a finite subgroup of O_n that maps lattice Λ to itself. The CSG G is a discrete subgroup of E_n, such that $L = G \cap T_n$ is a lattice group of the form (2.1) or, in other words, a discrete subgroup of the group of isometries of the 2-dimensional Euclidean space containing a lattice subgroup. An asymmetric unit is a subset of the primitive cell such that the whole $\mathbb{R}^n$ is filled when the CSG symmetry operations are applied.

The term CSG has two distinct meanings; one referring to an actual group and another to a group isomorphism class. We refer to the isomorphism classes of 2D CSGs as plane groups and 3D CSGs as space groups. 2D CSGs are classified into 17 isomorphism classes, and each class is assigned to one of the four maximal crystallographic point group conjugacy classes, referred to as the crystal systems. Similarly, 3D CSGs are classified into 32 crystal systems and further into 230 isomorphism classes.

Given an nD CSG $G \in \mathcal{G}$, where $\mathcal{G}$ is a CSG isomorphism class associated with G, and K a compact subset of $\mathbb{R}^n$, by CSG packing $\mathcal{K}_G$, we mean a collection of non-overlapping orbits of K with respect to the G-action on $\mathbb{R}^n$. Formally, $\mathcal{K}_G$ is a G-set defined as

$$(2.2a) \qquad \mathcal{K}_G = \bigcup_{g \in G} gK,$$

$$(2.2b) \qquad \mathrm{int}\,(g_iK) \cap \mathrm{int}\,(g_jK) = \emptyset, \quad \forall\, g_i,\, g_j \in G,\ g_i \neq g_j.$$

Every element g of the CSG G acting on some point $\mathbf{p}$ can be expressed as

$$(2.3) \qquad g\mathbf{p} = \mathbf{R}\mathbf{p} + \mathbf{a} + \mathbf{l},$$

where $\mathbf{R} \in O_n$, $\mathbf{a} \in T_\mathbf{a}$ such that

$$(2.4) \qquad \mathbf{a} = \alpha_1\mathbf{b}_1 + \ldots + \alpha_n\mathbf{b}_n$$

where $\alpha_i \in \mathbb{R}$, $0 \leq \alpha_i < 1$ and $\mathbf{b}_1, \ldots, \mathbf{b}_n$ lattice basis vectors, and $\mathbf{l} \in L$. Since $\mathcal{K}_G$ is a periodic system of sets it can be expressed in terms of (2.3) as $\mathbf{R}_iK + \mathbf{a}_i + \mathbf{l}_j$, with a

finite number of translation vectors $\mathbf{a}_i$ of form (2.4), rotations and rotation-inversions $R_i \in O_n$ for $i = 1, \ldots, N$, and lattice translation vectors $\mathbf{l}_j \in L$ for $j = 1, 2, \ldots$. Following from the formula for the packing density of the periodic system [64], the plane group packing density has a simple closed form expression

$$(2.5) \qquad\qquad \rho\left(\mathcal{K}_G\right) = \frac{N m(K)}{m(\overline{\Lambda})},$$

where N is the number of symmetry operations modulo lattice translations in a CSG G given by the pair $(\mathbf{R}, \mathbf{a})$ in (2.3), $\overline{\Lambda}$ is the primitive cell, and $m(\cdot)$ denotes the nD Jordan measure.

Finally, we can state the crystallographic packing problem. Given a CSG isomorphism class $\mathcal{G}$ and K, a compact subset of $\mathbb{R}^n$, the goal is to find the CSG packing $\mathcal{K}_{G_{\max}}$ with maximum density over the whole $\mathcal{G}$. Formally expressed, we aim at finding a $\mathcal{G}$-packing, such that

$$(2.6) \qquad\qquad \mathcal{K}_{G_{\max}} = \underset{\mathcal{K}_G \in \mathcal{G}}{\operatorname{argmax}}\, \rho\left(\mathcal{K}_G\right).$$

Since the volume of K and the number of symmetry operations N for a given nD CSG in (2.5) are constant, maximizing density is equivalent to minimizing the Jordan measure of the primitive cell associated with the crystal system of $\mathcal{G}$. The crystallographic packing problem can be restated as finding a $\mathcal{G}$-packing with minimal primitive cell volume

$$\mathcal{K}_{G_{\max}} = \underset{\mathcal{K}_G \in \mathcal{G}}{\operatorname{argmin}}\, m(\overline{\Lambda}).$$

We refer to the solution of (2.6) as the densest $\mathcal{G}$-packing. Here we search not only over the whole $\mathcal{G}$ but also over all rotations and translations of K, whose centroid lies in the asymmetric unit of G such that the resulting configuration is a CSG packing.

We consider the densest $\mathcal{G}$-packing as a nonlinear bounded constrained optimization problem. The Jordan measure of the primitive cell is computed as the determinant of lattice generators (2.1), a polynomial of degree n. The bounds are given by the space group's asymmetric unit, the range of rotational freedom of the set K, and bounds of the size and shape of the primitive cell given by the crystal system associated with $\mathcal{G}$. This means that the configuration space is compact, which implies that problem (2.6) has a solution. Linear and nonlinear constraints are given by the CSG's asymmetric unit and non-overlap condition (2.2b), respectively.

3. Entropic trust region. In the classic black box optimization setting, no requirements are imposed on the objective except that given any design parameters or query $\mathbf{x}$ from the set of possible configurations $\mathcal{X}$, the function value can be calculated. This kind of objective, which is here denoted by $\mathbf{F}$, is frequently referred to as a zero-order oracle due to the lack of accessibility to gradient information. In all of the following, we assume that $\mathbf{F}$ is a function between Borel measurable spaces $\mathcal{X}$ and $\mathcal{Y}$ with measures μ and ν, respectively, such that $\mu(\mathbf{F}(E)) = 0$ whenever $\nu(E) = 0$ for $E \subset \mathcal{X}$.

An approach for solving such optimization problems is stochastic relaxation [33], and it follows from the observation that optimizing $\mathbf{F}$ is equivalent to solving

$$(3.1) \qquad\qquad \tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}}\, J(\boldsymbol{\theta}),$$

with

$$J(\boldsymbol{\theta}) := E[\mathbf{F}|\boldsymbol{\theta}] = \int_{\mathcal{X}} \mathbf{F}(\mathbf{x}) dP(\boldsymbol{\theta}) \tag{3.2}$$

being the expected value of $\mathbf{F}$ under some probability measure $dP(\boldsymbol{\theta})$ from a parametric family of probability measures

$$S = \{dP(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^n\} \tag{3.3}$$

defined on some configuration space $\mathcal{X}$.

   Natural evolution strategies [85] solve (3.1) algorithmically using a non-euclidean trust region strategy where the candidate step $\delta\boldsymbol{\theta}$ is found by maximizing the first-order Taylor approximation of $J(\boldsymbol{\theta}+\delta\boldsymbol{\theta})$. The trust region radius is given by the square root of twice the second-order approximation of the Kullback-Leibler divergence from $P_{\boldsymbol{\theta}}$ to $P_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}$ in S. Specifically

$$\max_{\delta\boldsymbol{\theta}} J(\boldsymbol{\theta}^t) + \delta\boldsymbol{\theta}^{\mathsf{T}} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t) \tag{3.4a}$$

$$\text{s.t. } \sqrt{2D_{KL}\left(P_{\boldsymbol{\theta}} \mid\mid P_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}\right)} \approx \sqrt{\delta\boldsymbol{\theta}^{\mathsf{T}}\mathcal{I}_{\boldsymbol{\theta}}\delta\boldsymbol{\theta}} \leq \Delta^t, \tag{3.4b}$$

where $\nabla_{\boldsymbol{\theta}}$ denotes the euclidean gradient operator with respect to $\boldsymbol{\theta}$ coordinates, Δ^t is the trust region radius, and $\mathcal{I}_{\boldsymbol{\theta}}$ is the Fisher information matrix with elements

$$\mathcal{I}_{\boldsymbol{\theta} ij} = \int_{\mathcal{X}} \frac{\partial \ln\left(p(\boldsymbol{\theta})\right)}{\partial \theta_i} \frac{\partial \ln\left(p(\boldsymbol{\theta})\right)}{\partial \theta_j} dP(\boldsymbol{\theta}); \tag{3.5}$$

$p(\boldsymbol{\theta}) = \frac{dP(\boldsymbol{\theta})}{d\boldsymbol{\nu}}$ is the Radon-Nikodym derivative of $P(\boldsymbol{\theta})$ with respect to some reference measure $\boldsymbol{\nu}$ defined on $\mathcal{X}$, and the Kullback-Leibler divergence (KLD) from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}+\delta\boldsymbol{\theta}$ is given by

$$D_{KL}\left(P_{\boldsymbol{\theta}} \mid\mid P_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}\right) = \int_{\mathcal{X}} \ln\left(\frac{dP(\boldsymbol{\theta})}{dP(\boldsymbol{\theta} + \delta\boldsymbol{\theta})}\right) dP(\boldsymbol{\theta}) \tag{3.6}$$

where $P_{\boldsymbol{\theta}} \in S$ is parametrized by $\boldsymbol{\theta}$.

   Using Lagrange multipliers to solve (3.4) results in a trust region step size

$$\delta\boldsymbol{\theta}^t = \frac{\Delta^t \mathcal{I}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t)}{\sqrt{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t)^{\mathsf{T}} \mathcal{I}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t)}}. \tag{3.7}$$

   Expression (3.7) also has a differential geometric interpretation. By considering the statistical model S of (3.3), then

$$\text{grad}J(\boldsymbol{\theta}) = \frac{\mathcal{I}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})}{\sqrt{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})^{\mathsf{T}} \mathcal{I}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})}} \tag{3.8}$$

constitutes a geodesic vector field on some neighbourhood of $\boldsymbol{\theta}^t$ in the Riemannian manifold $(S, \mathcal{I}_{\boldsymbol{\theta}})$, with S being a statistical manifold [44] and with the Fisher information matrix $\mathcal{I}_{\boldsymbol{\theta}}$ used as the associated metric tensor. [1] refers to $\mathcal{I}_{\boldsymbol{\theta}}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ as the natural gradient, denoted by $\widetilde{\nabla} J(\boldsymbol{\theta})$.

   Kullbeck-Leibler divergence induces a local distance between $P_{\boldsymbol{\theta}}$ and $P_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}$ on S when $\delta\boldsymbol{\theta}$ is sufficiently small, given by $D_{KL}\left(P_{\boldsymbol{\theta}} \mid\mid P_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}\right) = \frac{1}{2}\delta\boldsymbol{\theta}^{\mathsf{T}}\mathcal{I}_{\boldsymbol{\theta}}\delta\boldsymbol{\theta}+\mathcal{O}\left(\parallel \delta\boldsymbol{\theta} \parallel^2\right)$, where

$\mathcal{O}(\| \delta\boldsymbol{\theta} \|^2)$ approaches zero as $\| \delta\boldsymbol{\theta} \|^2 \to 0$. The square of this distance is denoted by $ds^2 = \delta\boldsymbol{\theta}^\mathsf{T}\mathcal{I}_{\boldsymbol{\theta}}\delta\boldsymbol{\theta}$. Then, combining the natural gradient $\widetilde{\nabla}$ and the trust region radius constraints (3.4b) given by $ds \leq \Delta^t$, the update equations to solve (3.1) take the following form

$$(3.9) \qquad \boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \Delta^t \frac{\widetilde{\nabla} J(\boldsymbol{\theta}^t)}{\| \widetilde{\nabla} J(\boldsymbol{\theta}^t) \|_{\mathcal{I}_{\boldsymbol{\theta}}}},$$

where $\| \cdot \|_{\mathcal{I}_{\boldsymbol{\theta}}}$ is the norm associated with the inner product induced by the metric tensor $\mathcal{I}_{\boldsymbol{\theta}}$.

Model (3.4) can be considered a probabilistic equivalent of a euclidean trust region model [57]. Furthermore, as the search direction of the euclidean first-order trust region model coincides with the search direction of the line search method, (3.9) can be similarly viewed as a geodesic search on the statistical manifold S, where the search moves along a geodesic given by (3.8) with the step length given by Δ^t.

3.1. The extended multivariate von Mises distribution. Crystal lattice (2.1) induces a quotient space $\mathbb{R}^n/L$, where L is a lattice group (2.1). Since $\mathbb{R}^n/L$ is homeomorphic to the nD torus denoted by T^n, a natural choice for the statistical model S of (3.3) is to restrict the search to a family of probability distributions defined on T^n.

[49] defined a probability distribution with the support on a 2D unit flat torus and, using the sine submodel of the general bivariate von Mises model, introduced a probability distribution on an nD torus [50]. Following the general bivariate von Mises model, we extended the multivariate von Mises model to the family of distributions with the probability density function

$$(3.10) \quad f(\boldsymbol{\theta}|\boldsymbol{\mu},\boldsymbol{\kappa},\mathbf{D}) = \frac{1}{Z(\boldsymbol{\mu},\boldsymbol{\kappa},\mathbf{D})} \exp\left\{ \boldsymbol{\kappa}^\mathsf{T} c(\boldsymbol{\theta}-\boldsymbol{\mu}) + \frac{1}{2} \begin{bmatrix} c(\boldsymbol{\theta}-\boldsymbol{\mu}) \\ s(\boldsymbol{\theta}-\boldsymbol{\mu}) \end{bmatrix}^\mathsf{T} \mathbf{D} \begin{bmatrix} c(\boldsymbol{\theta}-\boldsymbol{\mu}) \\ s(\boldsymbol{\theta}-\boldsymbol{\mu}) \end{bmatrix} \right\}$$

where

$$c(\boldsymbol{\theta}-\boldsymbol{\mu}) = [cos(\theta_1-\mu_1),\ldots,cos(\theta_n-\mu_n)]^\mathsf{T},$$
$$s(\boldsymbol{\theta}-\boldsymbol{\mu}) = [sin(\theta_1-\mu_1),\ldots,sin(\theta_n-\mu_n)]^\mathsf{T},$$
$$0 \leq \theta_i, \mu_i \leq 2\pi, \quad 0 \leq \kappa_i,$$

and the $2n \times 2n$ real valued symmetric matrix that controls the cosine-sine interactions

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}^{cc} & \mathbf{D}^{cs} \\ \mathbf{D}^{cs\,\mathsf{T}} & \mathbf{D}^{ss} \end{bmatrix}$$

has all the diagonal elements d_{ii}^{cc}, d_{ii}^{ss}, d_{ii}^{cs} of its corresponding submatrices set to zero.

The explicit form of the normalizer $Z(\boldsymbol{\mu},\boldsymbol{\kappa},\mathbf{D})$ is known only in a few instances of the bivariate case [41, 70]. When an explicit evaluation is impossible, the standard approach to compute integrals is to use Monte–Carlo methods. For the exponential family statistical model used here, the Monte–Carlo estimates are discussed in Appendix A.

It has to be noted that the full multivariate von Mises model (3.10) is not identifiable and thus problematic to parameterize in the form of an exponential family of distributions since for $\boldsymbol{\kappa} = \mathbf{0}$ and any fixed $\mathbf{D}$, density functions (3.10) for any

$\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ are equal due to the vanishing of the term containing $\boldsymbol{\kappa}$. Nevertheless, when the concentration parameters $\boldsymbol{\kappa}$ are restricted to being strictly positive, the full multivariate von Mises model (3.10) can be rewritten using trigonometric identities in the following form

$$(3.11) \quad f(\boldsymbol{\theta}|\boldsymbol{\eta}, \mathbf{E}) = \exp\left\{ \begin{bmatrix} c(\boldsymbol{\theta}) \\ s(\boldsymbol{\theta}) \end{bmatrix}^{\mathsf{T}} \boldsymbol{\eta} + \mathrm{vec}\left(\begin{bmatrix} c(\boldsymbol{\theta}) \\ s(\boldsymbol{\theta}) \end{bmatrix} \begin{bmatrix} c(\boldsymbol{\theta}) \\ s(\boldsymbol{\theta}) \end{bmatrix}^{\mathsf{T}} \right)^{\mathsf{T}} \mathrm{vec}\left(\mathbf{E} \right) - \psi\left(\boldsymbol{\eta}, \mathbf{E} \right) \right\}$$

where $\mathrm{vec}(\cdot)$ denotes vectorization, $\psi\left(\cdot, \cdot\right)$ is the logarithm of the normalizing constant or log-partition function, and the canonical exponential family parameters $(\boldsymbol{\eta}, \mathbf{E})$ are given as

$$(3.12) \qquad\qquad \boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\kappa} \odot c(\boldsymbol{\mu}) \\ \boldsymbol{\kappa} \odot s(\boldsymbol{\mu}) \end{bmatrix}, \quad \mathbf{E} = \begin{bmatrix} \mathbf{E}^{cc} & \mathbf{E}^{cs} \\ \mathbf{E}^{cs\,\mathsf{T}} & \mathbf{E}^{ss} \end{bmatrix}.$$

where $\odot$ denotes Hadamard product. The submatrices can be expressed as follows

$$\mathbf{E}^{cc} = \tfrac{1}{2}\mathbf{D}^{cc} \odot c(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathsf{T}} - \tfrac{1}{2}\mathbf{D}^{cs} \odot c(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathsf{T}} + \tfrac{1}{2}\mathbf{D}^{ss} \odot s(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathsf{T}} - \tfrac{1}{2}\,\mathbf{D}^{cs\,\mathsf{T}} \odot s(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathsf{T}}$$

$$\mathbf{E}^{cs} = \tfrac{1}{2}\mathbf{D}^{cs} \odot c(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathsf{T}} + \tfrac{1}{2}\mathbf{D}^{cc} \odot c(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathsf{T}} - \tfrac{1}{2}\mathbf{D}^{ss} \odot s(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathsf{T}} - \tfrac{1}{2}\,\mathbf{D}^{cs\,\mathsf{T}} \odot s(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathsf{T}}$$

$$\mathbf{E}^{ss} = \tfrac{1}{2}\mathbf{D}^{ss} \odot c(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathsf{T}} + \tfrac{1}{2}\mathbf{D}^{cc} \odot s(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathsf{T}} + \tfrac{1}{2}\mathbf{D}^{cs} \odot s(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathsf{T}} + \tfrac{1}{2}\,\mathbf{D}^{cs\,\mathsf{T}} \odot c(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathsf{T}}.$$

The property that the reparameterisation of the extended multivariate von Mises model (3.11) is an exponential family probability distribution grants us a few beneficial properties besides the ability to use Monte–Carlo methods to estimate the natural gradients of Appendix A. For example, from the maximum entropy property of the exponential family, we have that (3.10) is the maximum entropy or minimum discrimination distribution when $(\boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{D})$ parameters are specified. Moreover exponential family provide the statistical model S (3.3) with a dually flat structure [2], a concept that is utilized in Subsection 3.3 and Subsection 3.4.

Additional details on the extended multivariate von Mises distribution, parameter transformations as well as implementation details for the Gibbs sampler used in the Monte–Carlo estimates are presented in Appendix B and subsections therein.

3.2. Adaptive selection quantile. Using the raw expected fitness (3.2) results in poor performance due to the sensitivity of the sample mean estimates of the expected fitness to extreme values [12]. The usual way to address this in evolutionary computation methods [58, 85] is to employ a rank-preserving transformation of the fitness function, referred to as the selection quantile [11]. The main advantage of using quantiles instead of raw fitnesses is that quantiles are highly robust measures of position. This section proposes a selection quantile fitness transformation inspired by the simulated annealing control parameter [43].

One particular selection quantile is of the form

$$(3.14) \qquad\qquad q\,\mathbf{1}_{\mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q}}}(\mathbf{x}) := \begin{cases} q & \text{if } \mathbf{F}(\mathbf{x}) \geq \mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q}}, \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathbf{F}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q}}$ is the $P(\tilde{\boldsymbol{\theta}})$ $(q-1)$th q-quantile of the fitness $\mathbf{F}$, using the indicator function $\mathbf{1}.(\cdot)$. Since the expected value of the indicator function (3.14) is equal to the

probability of observing $\mathbf{F}(\mathbf{x})$ being greater than $\mathbf{F}_{1-\frac{1}{q}}$ as

$$(3.15) \qquad J(\boldsymbol{\theta}) = \int_{\mathcal{X}} q \mathbf{1}_{\mathbf{F}_{1-\frac{1}{q}}^{\tilde{\theta}}}(\mathbf{x}) dP(\boldsymbol{\theta}) = q \int_{\mathcal{V}} dP(\boldsymbol{\theta}) = qP\left(\mathbf{F}(\mathbf{X}) \geq \mathbf{F}_{1-\frac{1}{q}}^{\tilde{\theta}} | \boldsymbol{\theta}\right)$$

where $\mathcal{V} = \{\mathbf{x} | \mathbf{F}(\mathbf{x}) \geq \mathbf{F}_{1-\frac{1}{q}}^{\tilde{\theta}}\}$ if we set $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$, then we have $qP\left(\mathbf{F}(\mathbf{X}) \geq \mathbf{F}_{1-\frac{1}{q}}^{\tilde{\theta}} | \tilde{\boldsymbol{\theta}}\right) = 1$, and

$$(3.16) \qquad J(y | \tilde{\boldsymbol{\theta}}) = \int_{\mathbf{F}_{1-\frac{1}{q}}^{\tilde{\theta}}}^{y} q dP \circ \mathbf{F}^{-1}(z)(\tilde{\boldsymbol{\theta}})$$

can be regarded as a probability distribution function of the random variable $Y = \mathbf{F}(\mathbf{X})$ truncated at $\mathbf{F}_{1-\frac{1}{q}}^{\tilde{\theta}}$ with the density function $qp(\mathbf{F}^{-1}(y) | \tilde{\boldsymbol{\theta}})$ where $p(\mathbf{x} | \tilde{\boldsymbol{\theta}})$ is the density function of the random variable $\mathbf{X}$.

Specifically, given a random vector $\mathbf{X}$ with the exponential family density function

$$(3.17) \qquad p(\mathbf{x} | \tilde{\boldsymbol{\theta}}) = \exp\left\{\tilde{\boldsymbol{\theta}}^{\mathsf{T}} \mathbf{t}(\mathbf{x}) - \psi\left(\tilde{\boldsymbol{\theta}}\right)\right\},$$

the density function of the random variable $Y = \mathbf{F}(\mathbf{X})$ truncated at $\mathbf{F}_{1-\frac{1}{q}}^{\tilde{\theta}}$ can be expressed as

$$(3.18) \qquad p(y | \tilde{\boldsymbol{\theta}}) = \exp\left\{\tilde{\boldsymbol{\theta}}^{\mathsf{T}} \mathbf{t}\left(\mathbf{F}^{-1}(y)\right) - \psi\left(\tilde{\boldsymbol{\theta}}\right) + \ln(q)\right\} w(y)$$

for some Borel measurable function w on $\mathcal{Y}$.

The $P(\tilde{\boldsymbol{\theta}})$ $(q-1)$th q-quantile of the fitness $\mathbf{F}$ is generally unknown and has to be estimated. Given N observations of $\mathbf{x}_i \sim P(\tilde{\boldsymbol{\theta}})$, the empirical quantile function is constructed by assigning the ranks

$$r_i^{\tilde{\theta}} := \left\{i \mid \mathbf{F}(\mathbf{x}_1) \leq \mathbf{F}(\mathbf{x}_2) \leq \ldots \leq \mathbf{F}(\mathbf{x}_i) \leq \ldots \leq \mathbf{F}(\mathbf{x}_{N-1}) \leq \mathbf{F}(\mathbf{x}_N) \mid \mathbf{x}_i \sim P(\tilde{\boldsymbol{\theta}})\right\}$$

to each fitness value. Then the $\mathbf{F}_{1-\frac{1}{q}}^{\tilde{\theta}}$ estimate $\hat{\mathbf{F}}_{1-\frac{1}{q}}^{\tilde{\theta}}$ is given by $\hat{\mathbf{F}}_{1-\frac{1}{q}}^{\tilde{\theta}} = \mathbf{F}(\mathbf{x}_{r_{\left\lfloor N-\frac{N}{q}\right\rfloor}^{\tilde{\theta}}})$, where $\lfloor \cdot \rfloor$ denotes rounding to the nearest lower integer.

This process can be repeated multiple times to estimate $\mathbf{F}_{1-\frac{1}{q}}^{\tilde{\theta}}$ by taking the maximum of the empirical quantile $\mathbf{F}(\mathbf{x}_{r_{\left\lfloor N-\frac{N}{q}\right\rfloor}^{\tilde{\theta}}})$ from all batch sampling iterations. The estimate can then take the form

$$\hat{\mathbf{F}}_{1-\frac{1}{q}}^{\tilde{\theta}} = \max_{i=1\ldots}\left[\mathbf{F}(\mathbf{x}_{r_{\left\lfloor N-\frac{N}{q}\right\rfloor}^{\tilde{\theta}}}^i)\right].$$

In practice, with the above, we simulate realizations from a truncated probability distribution of the distribution in (3.16), similarly to the simulated annealing method where the homogeneous algorithm [43] is a Metropolis-Hastings one for generating realizations from the Boltzmann distribution. Continuing with this analogy, the q-quantile can be considered equivalent to the temperature control parameter.

At the beginning of the search, it is beneficial to have a smaller q-quantile to maintain a more extensive profile of the overall optimization landscape. As the algorithm

progresses and we need the distribution to concentrate on samples with higher fitness values, higher values of q and a more localized search is preferred. We implement this using a time varying q_t-quantile by setting

$$(3.19) \qquad q_{t+1} = q_t \exp\{\beta t\}.$$

Consequently, for fixed $\tilde{\boldsymbol{\theta}}$, when t approaches infinity, the probability distribution (3.16) converges to a distribution where all the probability mass is concentrated on the extrema of the fitness F.

Following the fitness transformation (3.14), with $P(\boldsymbol{\theta})$ from the exponential family of distributions with the q-quantile fixed, the gradient of (3.15) at $\tilde{\boldsymbol{\theta}}$ takes the form

$$(3.20) \qquad \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\,|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = \boldsymbol{\mu}_{\mathbf{F}_{1-\frac{1}{q_t}}} - \boldsymbol{\mu},$$

where $\boldsymbol{\mu}_{\mathbf{F}_{1-\frac{1}{q_t}}}$ is the expectation parametrization of the truncated exponential probability distribution (3.18), given by

$$(3.21) \qquad \boldsymbol{\mu}_{\mathbf{F}_{1-\frac{1}{q_t}}} = \int_{\mathbf{F}^{\tilde{\theta}}_{1-\frac{1}{q_t}}}^{\infty} \mathbf{F}^{-1}(y)\exp\left\{\tilde{\boldsymbol{\theta}}^{\mathsf{T}}\mathbf{t}\left(\mathbf{F}^{-1}(y)\right) - \psi(\tilde{\boldsymbol{\theta}}) + \ln(q_t)\right\} d\mathbf{F}^{-1}(y),$$

and $\boldsymbol{\mu}$ is the expectation parametrization of exponential distribution with density function (3.17), equal to $\boldsymbol{\mu} = \int \mathbf{x}\exp\{\tilde{\boldsymbol{\theta}}^{\mathsf{T}}\mathbf{t}(\mathbf{x}) - \psi(\tilde{\boldsymbol{\theta}})\}dx$. Finally, the estimate of (3.20) is given by

$$(3.22) \qquad \nabla_{\boldsymbol{\theta}}\hat{J}(\boldsymbol{\theta})\,|_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}} = \hat{E}\left[\mathbf{t}(\mathbf{x})\,|\,\mathbf{F}(\mathbf{x}) \geq \hat{\mathbf{F}}^{\tilde{\boldsymbol{\theta}}}_{1-\frac{1}{q_t}}\right] - \hat{E}\left[\mathbf{t}(\mathbf{x})\right],$$

with $\hat{E}[\cdot]$ denoting the sample mean.

A caveat of simulated annealing with a fixed annealing schedule is the risk of the algorithm being trapped in a local maximum if the control parameter converges too fast. Therefore, the scheduling constant β in (3.19) has to be set optimally, which is usually done experimentally to counter this issue.

Another way to mitigate premature convergence caused by the control parameter (3.19) is by introducing self-adaptation into the control schedule where the selection quantile (3.14) is used in conjunction with the trust region method (3.9). For a fixed q_t, the algorithm performs a local search with the neighbourhood given by the q_t-quantile. By increasing q_t the neighbourhood becomes more localized, resulting in a decreased chance of escaping the attraction of local maxima. This behaviour is desirable in later stages when optima basins have already been selected. On the other hand, in instances when the trust region path reverses direction frequently, indicating a tendency toward moving to multiple local optima, it is beneficial to increase the search neighbourhood by decreasing q_t and provide the algorithm with less localized information for inference of various optima basins of the optimization landscape.

To assess the trajectory's current state, we compare the directions of three consecutive updates $\boldsymbol{\theta}^{t-2}, \boldsymbol{\theta}^{t-1}, \boldsymbol{\theta}^{t}$ by expressing their difference

$$\Delta\boldsymbol{\theta}^t = \boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1},$$
$$\Delta\boldsymbol{\theta}^{t-1} = \boldsymbol{\theta}^{t-1} - \boldsymbol{\theta}^{t-2},$$

as elements of the tangent space $T_{\boldsymbol{\theta}^{t-1}}S$ of a statistical manifold S at point $\boldsymbol{\theta}^{t-1}$. The angle α^t between vectors $\Delta\boldsymbol{\theta}^t$ and $\Delta\boldsymbol{\theta}^{t-1}$ induced by the scalar product $<\cdot,\cdot>_{\mathcal{I}_{\boldsymbol{\theta}^{t-1}}}$ with the Fisher metric tensor $\mathcal{I}_{\boldsymbol{\theta}^{t-1}}$ is given by

$$(3.23) \qquad \cos(\alpha^t) = \frac{<\Delta\boldsymbol{\theta}^t, \Delta\boldsymbol{\theta}^{t-1}>_{\mathcal{I}_{\boldsymbol{\theta}^{t-1}}}}{||\Delta\boldsymbol{\theta}^t||_{\mathcal{I}_{\boldsymbol{\theta}^{t-1}}}||\Delta\boldsymbol{\theta}^{t-1}||_{\mathcal{I}_{\boldsymbol{\theta}^{t-1}}}}.$$

By combining the quantile control parameter (3.19) and the direction of the current state of algorithm (3.23), we can use an adaptive selection quantile scheme as

$$(3.24) \qquad q_{t+1} = q_t \exp\left\{\beta\cos(\alpha^t)\right\} = q_0 \exp\left\{\beta\sum_{i=3}^{t}\cos(\alpha^i)\right\},$$

where q_0 is the initial value of the quantile control parameter.

The expression (3.20) has a clear geometric interpretation. $\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta})\mid_{\boldsymbol{\theta}=\tilde{\boldsymbol{\theta}}}$ can be regarded as a vector in the direction of $\boldsymbol{\mu}_{\mathbf{F}_{1-\frac{1}{q_t}}}$, that is, in the direction given by the (q_t-1)th q_t-quantile of the fitness. When the trust region steps (3.9) follow the same general direction, the adaptive selection quantile (3.24) increases, forcing the trust region to move toward higher fitness values. On the other hand, if $\cos(\alpha^t) < 0$, the algorithm backtracks and q_t is decreased, resulting in a broader search. In terms of evolutionary computation, the adaptive selection quantile models a type of time-varying selective pressure.

The proposed selection quantile can be used independently from the entropic trust region method, and its implementation details are provided in the form of pseudocode in Appendix C.

3.3. Entropic proximal maximization for exponential families. Proximal mappings are used in convex optimization to construct approximations of the objective function that have a smoothing effect and preserve the set of minimizers [62]. [9] showed equivalence between the mirror descent algorithm [56] and a modified proximal minimization algorithm by taking into account only first-order approximation of the objective function and considering a more general class of proximal operators [17]. Subsequently, [60] showed equivalence between the natural gradient descent method [1] and mirror descent. Following these results, we show that the entropic trust region method (3.9) for exponential family distributions is a special case of a proximal maximization where the projection term in the proximal operator is given by the KLD (3.6), resulting in a gradient ascent in dual coordinates, maximizing the likelihood and minimizing cross entropy of parameter update estimates.

Taking the first-order Taylor expansion of $J(\boldsymbol{\theta})$ in the neighbourhood of $\boldsymbol{\theta}^t$ with respect to exponential family statistical model S with canonical parameterization $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and expectation parameterization $\boldsymbol{\mu} \in \mathcal{M}$, the proximal method for problem (3.1) can be equivalently rewritten to

$$(3.25) \qquad \boldsymbol{\theta}^{t+1} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmax}}\left\{\boldsymbol{\theta}^{\mathsf{T}}\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta}^t) - \frac{1}{\epsilon}D_{KL}\left(P_{\boldsymbol{\theta}^t}\parallel P_{\boldsymbol{\theta}}\right)\right\},$$

with parameter $\epsilon > 0$ where the projection of $\boldsymbol{\theta}^t$ onto $\boldsymbol{\Theta}$ is given by the KLD. From the optimality condition of (3.25), we directly get the mirror ascent in the form

$$(3.26a) \qquad \boldsymbol{\mu}^{t+1} = \boldsymbol{\mu}^t + \epsilon\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta}^t),$$
$$(3.26b) \qquad \boldsymbol{\theta}^{t+1} = \nabla_{\boldsymbol{\mu}}\phi(\boldsymbol{\mu}^{t+1}),$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ are the dual canonical, and exponential parametrizations of the exponential family given by the Legendre transforms $\boldsymbol{\theta} = \nabla_{\boldsymbol{\mu}}\phi(\boldsymbol{\mu})$ and $\boldsymbol{\mu} = \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta})$ of the dual free energy functionals $\psi(\boldsymbol{\theta})$ and $\phi(\boldsymbol{\mu})$.

Differentiating $\boldsymbol{\mu}$ as a function of $\boldsymbol{\theta}$ with respect to time yields the time derivative

$$\frac{d\boldsymbol{\mu}(\boldsymbol{\theta})}{dt} = \nabla_{\boldsymbol{\theta}}^2 \psi(\boldsymbol{\theta}) \frac{d\boldsymbol{\theta}}{dt}.$$

For a sufficiently small dt and from the well-known fact that the Hessian of the free energy $\psi(\theta)$ is equal to the Fisher information matrix $\mathcal{I}_{\boldsymbol{\theta}}$ in $\boldsymbol{\theta}$ parameterization [8], the relationship between update step sizes in $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ parametrizations is given by

$$(3.27) \qquad \boldsymbol{\mu}^{t+1} - \boldsymbol{\mu}^t = \mathcal{I}_{\boldsymbol{\theta}^t}\left(\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\right).$$

Using (3.27), the mirror ascent update (3.26a) can be restated in the form of

$$(3.28) \qquad \boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \epsilon\mathcal{I}_{\boldsymbol{\theta}^t}^{-1}\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta}^t),$$

which is the natural gradient method [1] or information geometric optimization algorithm [58]. Furthermore, by setting

$$(3.29) \qquad \epsilon = \frac{\Delta^t}{\sqrt{\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta}^t)^{\mathsf{T}}\left(\mathcal{I}_{\boldsymbol{\theta}^t}\right)^{-1}\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta}^t)}},$$

we get the entropic trust region method of (3.9).

Statistically, the update $\boldsymbol{\mu}^{t+1}$ in (3.26a) is unknown, and we are working with the point estimate

$$(3.30) \qquad \hat{\boldsymbol{\mu}}^{t+1} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\mu}^t + \epsilon\mathbf{F}(\mathbf{X}_i)(\mathbf{X}_i - \boldsymbol{\mu}^t),$$

where $\mathbf{X}_i$ are independent identical exponential distributed family random vectors. This means that the update estimate $\hat{\boldsymbol{\mu}}^{t+1}$ is by itself a random vector from some parametric family of probability distributions, although generally not exponential family distributed. Nevertheless, using (3.30), the proximal map (3.25) can be rewritten as

$$\boldsymbol{\theta}^{t+1} = \operatorname*{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\{\boldsymbol{\theta}^{\mathsf{T}}\hat{\boldsymbol{\mu}}^{t+1} - \psi(\boldsymbol{\theta})\right\},$$

where $\ln p(\hat{\boldsymbol{\mu}}^{t+1}, \boldsymbol{\theta}) = \boldsymbol{\theta}^{\mathsf{T}}\hat{\boldsymbol{\mu}}^{t+1} - \psi(\boldsymbol{\theta})$ is the log-likelihood of the exponential family statistical model S given estimate (3.30). Thus, the mirror map projection step (3.26b) is given by the maximum likelihood estimate of $\boldsymbol{\theta}^{t+1}$.

Due to the duality of exponential families, we can rewrite the mirror ascent and thus the proximal map (3.25) in the expectation parameters providing us with dual trust region updates. Expressing the expected fitness (3.2) in terms of expectation parameters $\boldsymbol{\mu}$, yields

$$J(\boldsymbol{\theta}) = \int_{\mathbf{X}}\mathbf{F}(\mathbf{x})\exp\left\{\boldsymbol{\theta}^{\mathsf{T}}\mathbf{t}(\mathbf{x}) - \psi(\boldsymbol{\theta})\right\}dP_0(\mathbf{x}) =$$

$$= \int_{\mathbf{X}}\mathbf{F}(\mathbf{x})\exp\left\{\nabla_{\boldsymbol{\mu}}\phi(\boldsymbol{\mu})^{\mathsf{T}}(\mathbf{t}(\mathbf{x}) - \boldsymbol{\mu}) + \phi(\boldsymbol{\mu})\right\}dP_0(\mathbf{x}) = J(\boldsymbol{\mu}),$$

which, when differentiated with respect to $\boldsymbol{\mu}$, gives the $\boldsymbol{\theta}$ gradient of the expected fitness J as

$$(3.31) \qquad \nabla_{\boldsymbol{\mu}} J(\boldsymbol{\mu}) = \nabla_{\boldsymbol{\mu}}^2 \phi(\boldsymbol{\mu}) \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}).$$

Using (3.31) and the relationship between the Fisher information matrices in canonical and dual parametrizations and the Hessian of the Legendre dual ϕ given by

$$\mathcal{I}_{\boldsymbol{\theta}}^{-1} = \mathcal{I}_{\boldsymbol{\mu}} = \nabla_{\boldsymbol{\mu}}^2 \phi(\boldsymbol{\mu}),$$

produces the mirror ascent in the expectation parameters, which can be written as

$$(3.32a) \qquad \boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \epsilon \nabla_{\boldsymbol{\mu}} J(\boldsymbol{\mu}^t),$$

$$(3.32b) \qquad \boldsymbol{\mu}^{t+1} = \nabla_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta}^{t+1}),$$

or equivalently re-expressed as the proximal map maximization

$$(3.33) \qquad \boldsymbol{\mu}^{t+1} = \operatorname*{argmax}_{\boldsymbol{\mu} \in \mathcal{M}} \left\{ \boldsymbol{\mu}^{\mathsf{T}} \nabla_{\boldsymbol{\mu}} J(\boldsymbol{\mu}^t) - \frac{1}{\epsilon} D_{KL} \left(P_{\boldsymbol{\mu}} \,||\, P_{\boldsymbol{\mu}^t} \right) \right\}.$$

The preceding paragraphs make it clear that when the underlying statistical model is of the exponential family, the entropic trust region in (3.9) forms trajectories in the parameter spaces Θ and $\mathcal{M}$ related by the Bregman divergence derived from the free energy functional $\psi(\boldsymbol{\theta})$. In fact, these trajectories are given by a gradient flow of the expected fitness in dual coordinates, which will be further explored in the following section.

As in the case of the proximal map expressed in the canonical parametrization (3.25), the estimate of mirror ascent iteration (3.32a) defines the random vector

$$(3.34) \qquad \hat{\boldsymbol{\theta}}^{t+1} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\theta}^t + \epsilon \nabla_{\boldsymbol{\theta}}^2 \psi(\boldsymbol{\theta}^t)^{-1} \mathbf{F}(\mathbf{X}_i) \left(\mathbf{X}_i - \nabla_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta}^t) \right).$$

Then the proximal maximization (3.33) can be rewritten as

$$\boldsymbol{\mu}^{t+1} = \operatorname*{argmax}_{\boldsymbol{\mu} \in \mathcal{M}} \left\{ \boldsymbol{\mu}^{\mathsf{T}} \hat{\boldsymbol{\theta}}^{t+1} - \phi(\boldsymbol{\mu}) \right\},$$

and consequently as

$$\boldsymbol{\mu}^{t+1} = \operatorname*{argmin}_{\boldsymbol{\mu} \in \mathcal{M}} D_{KL} \left(P_{\nabla_{\boldsymbol{\mu}} \psi(\boldsymbol{\mu})} \,||\, Q_{\hat{\boldsymbol{\theta}}^{t+1}} \right),$$

assuming that $\hat{\boldsymbol{\theta}}^{t+1}$ belongs to a domain of the log-partition function $\psi(\cdot)$ and using the convex conjugate $\psi(\boldsymbol{\theta}) = \boldsymbol{\theta}^{\mathsf{T}} \boldsymbol{\mu} - \phi(\boldsymbol{\mu})$, where $Q_{\hat{\boldsymbol{\theta}}^{t+1}}$ is the probability distribution of (3.34). Thus, the mirror ascent projection (3.32b) is the minimum cross entropy distribution from S, given the prior distribution of (3.34) [68].

3.4. Geometry of the exponential family quantile-based entropic trust region method. Subsection 3.3 introduced a relationship between the entropic trust region method and entropic proximal mappings for the particular case when the search is performed on an exponential family parametric statistical model. When the quantile re-expression of the fitness of subsection 3.2 is added, the dual gradient flows

associated with the mirror ascents (3.26) or alternatively (3.32) result in a geometrical interpretation relating evolutionary strategies, simulated annealing method and recurrent neural computing as instances of more general graphical interaction models [21]. This section presents the exponential family quantile-based entropic trust region as an iterative solution to a minimax problem of the KLD between two hypersurfaces of an ambient statistical manifold.

We consider a general exponential family

$$(3.35) \qquad S^e := \{dP^e(\boldsymbol{\theta}) = \exp\{\boldsymbol{\theta}^\mathsf{T}\mathbf{t}(\mathbf{x}) - \psi(\boldsymbol{\theta})\}dP_0(\mathbf{x})\},$$

where dP^e is absolutely continuous with respect to some reference measure $dP_0(\mathbf{x})$ and $\psi(\boldsymbol{\theta}) = \ln \int \exp\{\boldsymbol{\theta}^\mathsf{T}\mathbf{t}(\mathbf{x})\}dP_0(\mathbf{x})$. Given random variable $Y = \mathbf{F}(\mathbf{X})$, where $\mathbf{F}$ is as in section 3, S^e can be rewritten to

$$(3.36) \qquad S^e = \{dP^e(\boldsymbol{\theta})| \boldsymbol{\theta} \in \Theta^e \subset \mathbb{R}^n\},$$

where $dP^e(\boldsymbol{\theta}) = \exp\{\boldsymbol{\theta}^\mathsf{T}\mathbf{t}\left(\mathbf{F}^{-1}(y)\right) - \psi(\boldsymbol{\theta})\}dP_0 \circ \mathbf{F}^{-1}(y)$ and $P_0 \circ \mathbf{F}^{-1}$ denotes the image of measure P_0 under the mapping $\mathbf{F}$, and a family of probability distributions derived from (3.36), introduced in subsection 3.2, by truncating S^e at the $P(\tilde{\boldsymbol{\theta}})$ $(q\text{–}1)$th q-quantile of fitness $\mathbf{F}$, as

$$(3.37) \qquad S^q = \{dP^q(\boldsymbol{\theta})| \boldsymbol{\theta} \in \Theta^q \subset \mathbb{R}^n\},$$

where

$$dP^q = \begin{cases} \exp\left\{\boldsymbol{\theta}^\mathsf{T}\mathbf{t}\left(\mathbf{F}^{-1}(y)\right) - \psi(\boldsymbol{\theta}) + \ln(q)\right\}dP_0 \circ \mathbf{F}^{-1}(y) & y \geq \mathbf{F}^{\boldsymbol{\theta}}_{1-\frac{1}{q}} \\ 0 & \text{otherwise} \end{cases}$$

and $\frac{1}{q} = \int_{\mathbf{F}^{\boldsymbol{\theta}}_{1-\frac{1}{q}}}^{\infty} \exp\{\boldsymbol{\theta}^\mathsf{T}\mathbf{t}\left(\mathbf{F}^{-1}(y)\right) - \psi(\boldsymbol{\theta})\}dP_0 \circ \mathbf{F}^{-1}(y)$ with truncation parameter $\gamma = \mathbf{F}^{\boldsymbol{\theta}}_{1-\frac{1}{q}}$.

For a fixed truncation parameter γ, the parametric family of probability distributions S^q is that of a regular exponential family derived from the S^e. On the other hand, if we consider $q \in [1; \infty)$ as another parameter, we can think of S^q as the leaves of a foliation [45] of some statistical manifold and set a new family of probability distributions as

$$(3.38) \qquad S = \bigcup_{q \in [1;\infty)} S^q$$

with the parametric space $\Theta \times [1; \infty) \subset \mathbb{R}^{n+1}$. Since the truncation parameter in (3.37) is exactly determined by $\boldsymbol{\theta}$ and q, S retains the dual structure given by the Bregman divergence [14] induced by the convex function

$$(3.39) \qquad \psi(\boldsymbol{\theta}^q) = \psi(\boldsymbol{\theta}) - \ln(q),$$

with the parametrization $\boldsymbol{\theta}^q = [\boldsymbol{\theta}, q]$. Moreover, since there is a bijection between regular exponential families and regular Bregman divergences [7], S is a regular exponential family. Consequently, the dual expectation parametrization is given by $\boldsymbol{\mu}^q = \left[\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}^q), -\frac{1}{q}\right]$ and the Fisher metric tensor by $\mathcal{I}_{\boldsymbol{\theta}^q} = \begin{bmatrix} \nabla^2_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}^q) & \mathbf{0} \\ \mathbf{0} & q^{-2} \end{bmatrix}$.

In the following, we show that the entropic proximal maximization algorithm of subsection 3.3 with the selection quantile fitness transformation (3.14) constitutes dual

gradient flows in S, maximizing the KLD from S^{q^*} (3.37) to S^1 (3.36) and minimizing this divergence from S^1 (3.37) to S^{q^*} (3.36). In mathematical terms, we wish to solve the following max-min problem,

$$(3.40) \qquad \max_{P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}} \min_{P_{\boldsymbol{\theta}^1} \in S^1} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \| P_{\boldsymbol{\theta}^1}).$$

By fixing $P_{\boldsymbol{\theta}^{q^*}}$, we can minimize $D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \| P_{\boldsymbol{\theta}^1})$ with respect to $P_{\boldsymbol{\theta}^1}$. Since the KLD from $P_{\boldsymbol{\theta}^{q^*}}$ to $P_{\boldsymbol{\theta}^e}$ equals the Bregman divergence derived from the convex function (3.39), we have

$$(3.41) \quad D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \| P_{\boldsymbol{\theta}^1}) = D_\psi(\boldsymbol{\theta}^1, \boldsymbol{\theta}^{q^*}) =$$
$$= \psi(\boldsymbol{\theta}^q)\,|_{\boldsymbol{\theta}^1} - \psi(\boldsymbol{\theta}^q)\,|_{\boldsymbol{\theta}^{q^*}} - \nabla_{\boldsymbol{\theta}^q} \psi(\boldsymbol{\theta}^q)^{\mathsf{T}}\,|_{\boldsymbol{\theta}^{q^*}} (\boldsymbol{\theta}^1 - \boldsymbol{\theta}^{q^*}).$$

Differentiating with respect to $\boldsymbol{\theta}^1$, the Riemannian gradient of (3.41) takes the form

$$(3.42) \quad \mathrm{grad} D_\psi(\boldsymbol{\theta}^1, \boldsymbol{\theta}^{q^*}) = \mathcal{I}_{\boldsymbol{\theta}^1}^{-1} \left(\nabla_{\boldsymbol{\theta}^q} \psi(\boldsymbol{\theta}^q)\,|_{\boldsymbol{\theta}^1} - \nabla_{\boldsymbol{\theta}^q} \psi(\boldsymbol{\theta}^q)\,|_{\boldsymbol{\theta}^{q^*}} \right) = \; = \mathcal{I}_{\boldsymbol{\theta}^1}^{-1}(\boldsymbol{\mu}^1 - \boldsymbol{\mu}^{q^*}),$$

and the gradient descent update equations to solve the minimization part of (3.40) then takes the following form,

$$(3.43) \qquad \boldsymbol{\theta}^{1^{t+1}} = \boldsymbol{\theta}^{1^t} + \epsilon \mathcal{I}_{\boldsymbol{\theta}^{1^t}}^{-1}(\boldsymbol{\mu}^{q^{*t}} - \boldsymbol{\mu}^{1^t}).$$

On the other hand, for a fixed $P_{\boldsymbol{\theta}^1}$, we can minimize $D_{KL}(P_{\boldsymbol{\theta}^q} \| P_{\boldsymbol{\theta}^1})$ with respect to $\boldsymbol{\theta}^q$. First, we express the KLD from $P_{\boldsymbol{\theta}^{q^*}}$ to $P_{\boldsymbol{\theta}^1}$ in terms of the Bregman divergence derived from the Legendre dual of $\psi(\boldsymbol{\theta}^q)$, given by $\phi(\boldsymbol{\mu}^q) = \boldsymbol{\theta}^{q\mathsf{T}} \boldsymbol{\mu}^q - \psi(\boldsymbol{\theta}^q)$, to the following form

$$(3.44) \quad D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \| P_{\boldsymbol{\theta}^1}) = D_\phi(\boldsymbol{\mu}^{q^*}, \boldsymbol{\mu}^1) =$$
$$= \phi(\boldsymbol{\mu}^q)\,|_{\boldsymbol{\mu}^{q^*}} - \phi(\boldsymbol{\mu}^q)\,|_{\boldsymbol{\mu}^1} - \nabla_{\boldsymbol{\mu}^q} \phi(\boldsymbol{\mu}^q)^{\mathsf{T}}\,|_{\boldsymbol{\mu}^{q^*}} (\boldsymbol{\mu}^{q^*} - \boldsymbol{\mu}^1)$$

and then by differentiating (3.44) with respect to $\boldsymbol{\mu}^{q^*}$, the Riemannian gradient is expressed as

$$(3.45) \quad \mathrm{grad} D_\phi(\boldsymbol{\mu}^{q^*}, \boldsymbol{\mu}^1) = \mathcal{I}_{\boldsymbol{\mu}^{q^*}}^{-1} \left(\nabla_{\boldsymbol{\mu}^q} \phi(\boldsymbol{\mu}^q)\,|_{\boldsymbol{\mu}^{q^*}} - \nabla_{\boldsymbol{\mu}^q} \phi(\boldsymbol{\mu}^q)\,|_{\boldsymbol{\mu}^1} \right) = \mathcal{I}_{\boldsymbol{\mu}^{q^*}}^{-1}(\boldsymbol{\theta}^{q^*} - \boldsymbol{\theta}^1).$$

Then, the gradient ascent update equations to solve the maximization part of (3.40) are

$$(3.46) \qquad \boldsymbol{\mu}^{q^{*t+1}} = \boldsymbol{\mu}^{q^{*t}} + \epsilon \mathcal{I}_{\boldsymbol{\mu}^{q^{*t}}}^{-1}(\boldsymbol{\theta}^{q^{*t}} - \boldsymbol{\theta}^{1^t}).$$

For a fixed time t, the trust region search is being performed on some neighbourhood of $P_{\boldsymbol{\theta}^t}$ independently from q^t since q^t is updated between trust region updates by (3.24). Considering S^{q^*} and S^1 as embeddings of S^e (3.35) in S (3.38), we recover S^e from S via the projection $[\theta_i^1]_{i=1,\dots,n} \to \boldsymbol{\theta}$ in the $\boldsymbol{\theta}$-coordinates for $\boldsymbol{\theta}^1 \in S^1$. For $\boldsymbol{\mu}^{q^*} \in S^{q^*}$, the $\boldsymbol{\mu}$-coordinates of S^e are such that the following expression holds, $\boldsymbol{\mu} = \frac{1}{q^*}[\mu_i^{q^*}]_{i=1,\dots,n} + \frac{q^*-1}{q^*} \boldsymbol{\mu}_{\mathbf{F}^C_{1-\frac{1}{q^*}}}$, where

$$\boldsymbol{\mu}_{\mathbf{F}^C_{1-\frac{1}{q^*}}} = \int_{-\infty}^{\mathbf{F}_{1-\frac{1}{q^*}}} \mathbf{F}^{-1}(y) \exp\left\{ \boldsymbol{\theta}^{\mathsf{T}} \mathbf{t}\left(\mathbf{F}^{-1}(y)\right) - \psi(\boldsymbol{\theta}) + \ln\left(\frac{q^*}{q^*-1}\right) \right\} dP_0 \circ \mathbf{F}^{-1}(y)$$

is the expected value of the complementary distribution to the truncated exponential distribution parameterized by $\boldsymbol{\theta}^{q^*}$ in (3.37). From the dual gradient updates (3.43) and (3.46) we then get the proximal maximization map in dual coordinates (3.25) and (3.33) for the expected fitness $J(\boldsymbol{\theta})$ given by the selection quantile (3.14), and consequently the $\boldsymbol{\theta}$-coordinate trust region update (3.9) by setting ϵ as in (3.29).

[5] analyzed the infomax principle proposed by [47] as a universal rule to train artificial neural networks based on observations in biological neural networks. The infomax principle states that network weights should be set in such a way that the mutual information between input and output is maximized. Later, [6] extended this work to maximizers of multi-information, a generalization of mutual information, previously defined in [82] as total correlation and later proposed by [72] as a measure of stochastic dependence in complex systems. Under this perspective, the dual gradient flow induced by (3.42) and (3.45) can be interpreted as an instance of the generalized infomax principle.

To obtain a better view of this, we can first observe that the maximum entropy estimate gives the solution to the minimizing part of (3.40). Let

$$\hat{\boldsymbol{\theta}}^1 = \underset{P_{\boldsymbol{\theta}^1} \in S^1}{\operatorname{argmin}} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \parallel P_{\boldsymbol{\theta}^1})$$

denote this estimate for some $P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}$ and $P_{\hat{\boldsymbol{\theta}}^1_s} \in S^1$ a split model corresponding to the maximum entropy estimate $\hat{\boldsymbol{\theta}}^1$ by setting all interaction parameters to zero. In our case, this means setting all the elements of the $\mathbf{E}$ matrix (3.12) in the exponential reparametrization of the full multivariate von Mises model (3.11) to zero. Since the Bregman divergence associated with the free energy (3.39) induces a dually affine flat structure on S (3.38), the information projection theorems and the generalized Pythagorean relation hold [2]. That is, $\hat{\boldsymbol{\theta}}^1$ is given by an orthogonal geodesic projection of $P_{\boldsymbol{\theta}^{q^*}}$ onto S^1, meaning that the dual geodesic connecting $P_{\boldsymbol{\theta}^{q^*}}$ and $P_{\hat{\boldsymbol{\theta}}^1}$ and the geodesic connecting $P_{\hat{\boldsymbol{\theta}}^1}$ and $P_{\hat{\boldsymbol{\theta}}^1_s}$ are orthogonal intersecting at $P_{\hat{\boldsymbol{\theta}}^1}$. Thus, the KLD from $P_{\boldsymbol{\theta}^{q^*}}$ to $P_{\hat{\boldsymbol{\theta}}^1_s}$ can be decomposed into $D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \parallel P_{\hat{\boldsymbol{\theta}}^1_s}) = D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \parallel P_{\hat{\boldsymbol{\theta}}^1}) + D_{KL}(P_{\hat{\boldsymbol{\theta}}^1} \parallel P_{\hat{\boldsymbol{\theta}}^1_s})$ or alternatively in the form $D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \parallel P_{\hat{\boldsymbol{\theta}}^1}) = D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \parallel P_{\hat{\boldsymbol{\theta}}^1_s}) - D_{KL}(P_{\hat{\boldsymbol{\theta}}^1} \parallel P_{\hat{\boldsymbol{\theta}}^1_s})$. Subsequently, for a fixed $P_{\hat{\boldsymbol{\theta}}^1}$ solving $\max_{P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \parallel P_{\hat{\boldsymbol{\theta}}^1})$ in (3.40) is equivalent to solving

$$(3.47) \qquad \max_{P_{\boldsymbol{\theta}^{q^*}} \in S^{q^*}} D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \parallel P_{\hat{\boldsymbol{\theta}}^1_s}).$$

Multi-information is defined as the KLD from the joint probability distribution $p(x_1, \ldots, x_n)$ to the product of the marginals $p(x_1) * \ldots * p(x_n)$ of a random vector $\mathbf{X} = (X_1, \ldots, X_n)$. It can state that the components of the random vector $\mathbf{X}$ are independent; that is, $p(x_1, \ldots, x_n) = p(x_1) * \ldots * p(x_n)$ if and only if the KLD or the multi-information vanishes. Given a split model, that is a model without any interactions between state variables, parametrized by $\hat{\boldsymbol{\theta}}^1_S$ of form (3.36), the components of the random vector associated with the split model are independent and the joint probability measure $dP(\hat{\boldsymbol{\theta}}^1_s)$ can be decomposed into a product of marginal measures

$$dP(\hat{\boldsymbol{\theta}}^1_s) = \left[\prod_{i=1}^n \exp\left\{ \left(\hat{\theta}_s^{\,1}\right)_i t_i \left((\mathbf{F}^{-1}(y))_i\right) \right\} \right] \exp\left\{ -\psi\left(\hat{\boldsymbol{\theta}}^1_s\right) \right\} dP_0 \circ \mathbf{F}^{-1}(y),$$

where $t_i(x_i)$ is the ith component of the sufficient statistic $\mathbf{t}(\mathbf{x}) = (t_1(x_1), \ldots, t_n(x_n))$.

Furthermore, the KLD in (3.47) can be rewritten as

$$D_{KL}(P_{\boldsymbol{\theta}^{q^*}} \parallel P_{\hat{\boldsymbol{\theta}}_s^1}) = \int\limits_{\left\{\mathbf{F}_{1-\frac{1}{q^*}} \leq \mathbf{F}(\mathbf{x})\right\}} \ln \frac{dP(\boldsymbol{\theta}^{q^*})}{dP(\hat{\boldsymbol{\theta}}_s^1)} dP(\boldsymbol{\theta}^{q^*}).$$

due to $dP(\boldsymbol{\theta}^{q^*}) = 0$ for $\mathbf{F}_{1-\frac{1}{q^*}} > \mathbf{F}(\mathbf{x})$. Since q^* is constant for a given iteration, the dual gradient ascent (3.43) and (3.46) and the related entropic proximal ascent in dual coordinates (3.25) and (3.33) can be interpreted as moving towards maximizing stochastic dependence. In other words, they are increasing the amount of information shared among elements of the random vector $\mathbf{X}$ due to the maximization of multi-information from the truncated exponential family S^{q^*} defined by the selection quantile (3.14) to the exponential family S^1 with full support.

3.5. Refining solutions. Stochastic optimization methods are generally well suited for problems where approximate solutions are sufficient. Overall, it is practically impossible to sample the exact solution, even when, in theory, an algorithm converges to a delta distribution concentrated on the solutions of the optimization problem. The entropic trust region (3.9) is no exception. To attain higher accuracy, we use the following procedure to refine the solution attained after the initial optimization by reducing the configuration space to a smaller region.

An $\boldsymbol{\epsilon}^r$ neighbourhood is created around the current best solution $\mathbf{x}^{\text{best}}$. The optimal so far solution at the rth run is recorded as $\mathbf{x}^{\text{best}} = \{\mathbf{x}^s | \mathbf{F}(\mathbf{x}^s) \geq \mathbf{F}(\mathbf{x}_i), \ i = 0, 1 \ldots, r\}$, with $\boldsymbol{\epsilon}^r$ set to

$$(3.48) \qquad\qquad \boldsymbol{\epsilon}^r = \left(\frac{1}{c_\epsilon}\right)^r (\mathbf{u} - \mathbf{l}),$$

where $c_\epsilon > 1$ is the rate of exponential decay, $r = 1, 2, \ldots$ denotes the current run of the refining process and $\mathbf{u}$, $\mathbf{l}$ are the initial upper and lower bounds, given by the optimization problem.

Based on $\boldsymbol{\epsilon}^r$, we reduce the two bounds $\mathbf{u}_\epsilon$, $\mathbf{l}_\epsilon$ in the next run by setting

$$(3.49) \qquad\qquad \begin{aligned} \mathbf{u}^\epsilon &= \min\left(\mathbf{x}^{\text{best}} + \boldsymbol{\epsilon}^r, \mathbf{u}\right), \\ \mathbf{l}^\epsilon &= \max\left(\mathbf{x}^{\text{best}} - \boldsymbol{\epsilon}^r, \mathbf{l}\right), \end{aligned}$$

taken along respective optimization variables and proceed with the next iteration of the algorithm set to the new upper and lower bounds $\mathbf{u}_\epsilon$, $\mathbf{l}_\epsilon$. We iterate this process until the desired accuracy is achieved. For example, in the setting of CSG packings, accuracy can be measured by the minimal euclidean distance between sets gK of a G-packing $\mathcal{K}_G$ in (2.2). The algorithm concentrates on the expected optimum with higher precision with each run. This refining process can be interpreted as a variant of a restart strategy.

4. Experimentations: A case study for dense plane group packings of convex polygons. In this section, we analyze the behaviour and performance of the introduced entropic trust region on the problem of packing convex polygons in 2D CSGs. In section 2, we introduced these symmetry groups, also referred to as plane groups, and defined a subproblem of the general packing problem [64]. Specifically, we present an experimental examination for the case study of the densest $p2$-packings of a regular octagon and results from implementations of our packing algorithm to
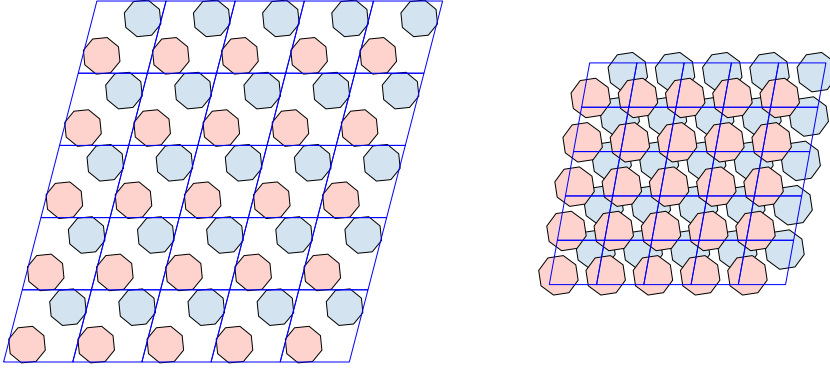
Figure 2: Visualization of 25 cells of two $p2$ configurations of a regular octagon from the initial sampling in Figure 3. (Left) Feasible solution of a $p2$-packing with density $\rho\left(\mathcal{K}_{p2}\right) = 0.413705837593271$. (Right) Unfeasible solution due to overlapping.
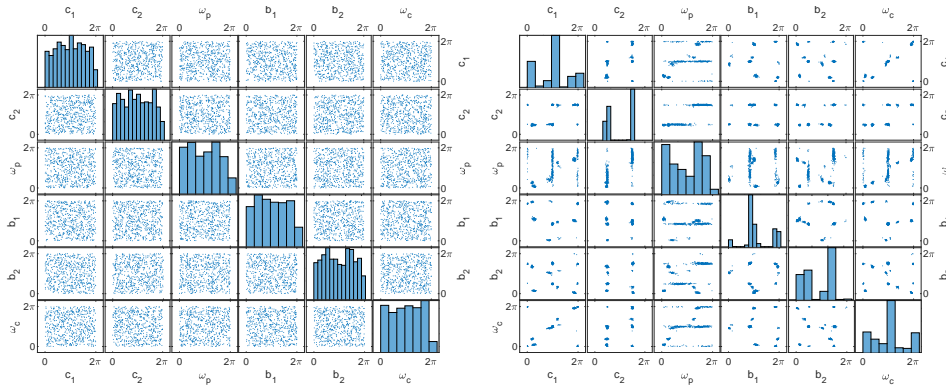

Figure 3: 2D projections along coordinate axes and histograms of univariate marginals corresponding to the respective optimization variables of 600 realizations of the exponential reparametrization of the full multivariate von Mises distribution. (Left) Initial distribution. (Right) Output distribution.

various plane group packings of convex polygons in instances for which the optimal solutions are known Table 1. More experimental results are presented in the supplementary material. That is additional details related to the behaviour of entropic trust region in $p2$-packing of regular octagons test case (examination of distribution parameter trajectories Appendix E and dependence of stability on the sampling pool size Appendix F) and a more detailed presentation of results from additional plane group packings (Appendix G). Technical details concerning the implementation of the entropic trust region method to the densest plane group packings are discussed in Appendix D [1].

Here, we chose the problem of $p2$-packing of convex octagons to showcase the behaviour and performance of the algorithm for several reasons. First, the general optimal packing of octagons is known. Second, the problem, as we have defined it in

---

[1] Matlab source code, instructions to use and examples are available at http://www.milotorda.net/wp-content/uploads/packing.zip/.
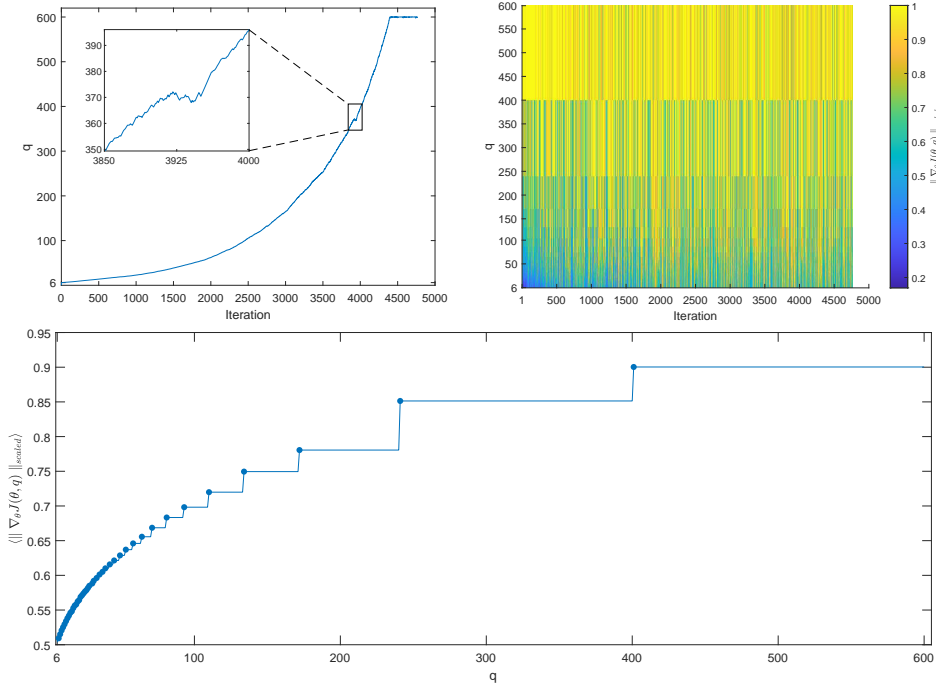
Figure 4: (Top left) Evolution of the adaptive selection quantile q. (Top right) Relationship between selection quantile q and the scaled expected fitness gradient (4.1) at every iteration of the entropic trust region run, and (Bottom) the scaled expected fitness gradient averaged through iterations where markers denote changes in the value of $<\| \nabla_\theta J(\theta, q) \|_{scaled}>$.

section 2, has at least 64 global maxima due to the 8-fold rotational symmetry of the octagon. The way the plane groups are constructed and the initial boundary values make this case more challenging than the other test cases.

The $p2$ plane group is a semi-direct product of a point group and a lattice group introduced in section 2. The point group is given by two symmetry operations, that is the identity operation ($\mathbf{R} = \mathbf{I}$) and a rotation by 180°. The lattice group, given by the oblique crystal system, has three degrees of freedom, namely the length of the lattice group generators or basic vectors $\mathbf{b}_1$, $\mathbf{b}_2$ and an angle ω_c between them, constituting part of the design variables to optimize. Additional degrees of freedom of the optimization configuration space are given by the fractional coordinates (a coordinate system where basis vectors are defined by the lattice basic vectors (2.1)) of the position of the centroid of the octagon in the asymmetric unit $\mathbf{c}_1$, $\mathbf{c}_2$ and an angle of rotation ω_p of the octagon. Altogether, we have six optimization variables for this specific test case.

The optimal general packing of centrally symmetric convex polygons is lattice packing [63]. Since the regular octagon is a polygon with central symmetry, the packing is known with the density of $\rho_{\text{opt}} = \frac{4+4\sqrt{2}}{5+4\sqrt{2}} \approx 0.9061636$.

The optimization boundaries are set as follows. From the asymmetric unit restrictions of the $p2$ group, octagon centroid fractional coordinates $c_1 \in (0, 1]$ and $c_2 \in \left(0, \frac{1}{2}\right]$, angle of rotation of the octagon $\omega_p \in (0, 2\pi]$, lengths of the lattice generators $b_i \in [0, 2d]$ for $i = 1, 2$ where d denotes the diameter of the octagon's circumcircle,
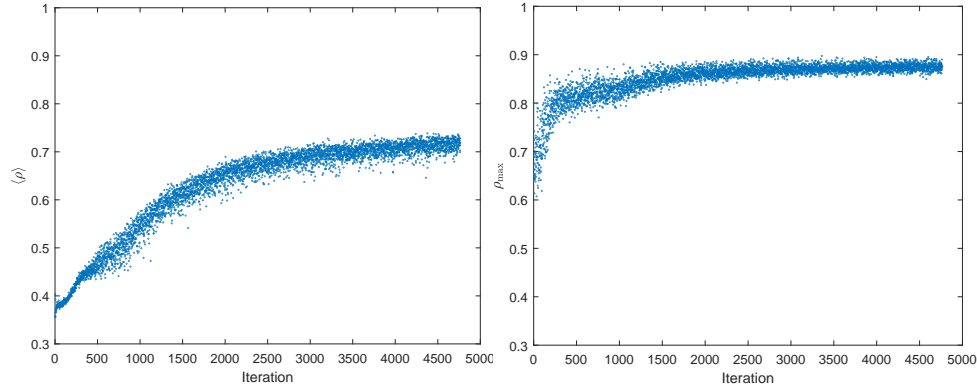
Figure 5: Evolution of (left) arithmetic mean and (right) maximum packing density of 600 solutions generated in each iteration.
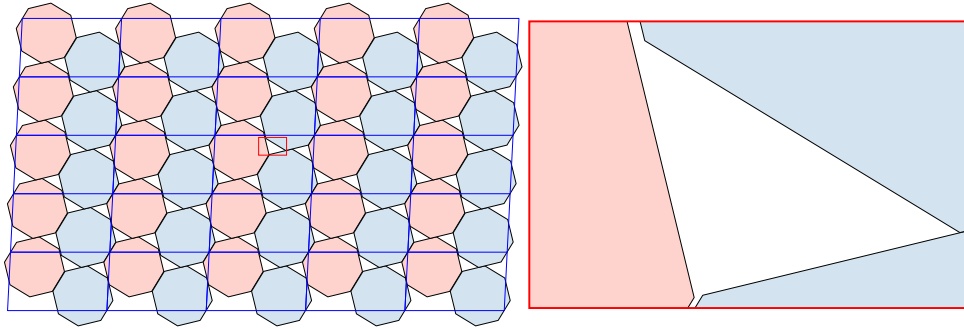

Figure 6: Best solution found during the initial optimization run. (Left) 25 cells drawn from the whole packing configuration. (Right) An enlargement of the configuration segment marked by a red rectangle.

and angle between lattice generators $\omega_c \in \left[0, \frac{\pi}{2}\right]$.

At the beginning of each optimization run all of the parameters of the exponential reparametrization of the full multivariate von Mises distribution $(\boldsymbol{\eta}, \mathbf{E})$ in (3.11) are set to zero, and 600 realizations from this distribution are generated. Effectively, at the start of each optimization run, we are sampling from a uniform distribution on a 6D flat torus, providing a homogeneous and unbiased first exploration of the optimization landscape. Figure 3 illustrates 2D projections along the respective coordinate axes of 600 samples generated in this way. Two candidate solutions from these initial 600 samples are shown in Figure 2.

Since there is no correspondence between the full multivariate von Mises distribution (3.10) and its exponential form (3.11) for $(\boldsymbol{\eta}, \mathbf{E}) = \mathbf{0}$, the initial trust region update is done using (3.9) for the exponential von Mises distribution (3.11). By doing so, we remove the dependency of the optimization on the initial position that burdens many optimization methods. After the initial step, all updates are performed in the $(\boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{D})$ parametrization (3.10).

During execution, the q parameter of the adaptive selection quantile (3.24) gradually increases, as shown in Figure 4 and intended by our design. The starting value of q is 6, meaning that at the beginning, 99% of the 600 samples are taken into account
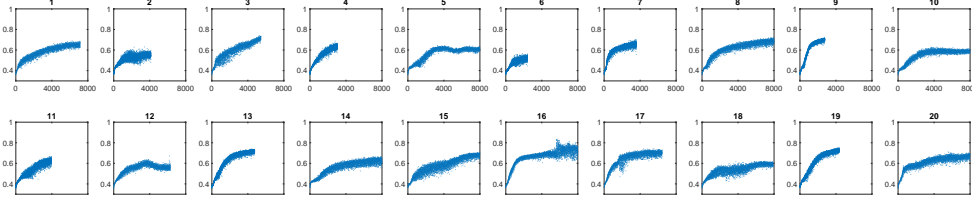
Figure 7: Evolution of the average density $\langle \rho \rangle$ in 20 optimization runs.
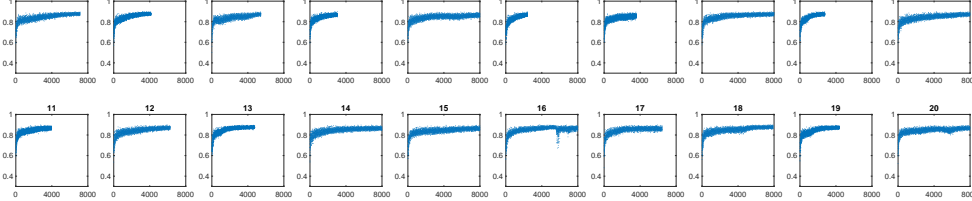
Figure 8: Evolution of maximum density $\rho_{\max}$ in 20 optimization runs.

for the truncated exponential distribution expected value $\boldsymbol{\mu}_{\mathbf{F}_{1-\frac{1}{q_t}}}$ (3.21). After about $4,000$ iterations, q reaches 600, which is the overall number of samples used in each iteration. Meaning, that only a single realization from the sampling distribution with the highest fitness is assigned to $\boldsymbol{\mu}_{\mathbf{F}_{1-\frac{1}{q_t}}}$, and the algorithm moves in the direction of this point. An interesting observation from the evolution of q is that the directions between the vectors in the tangent space $T_{\boldsymbol{\theta}^{t-1}}S_{\boldsymbol{\theta}}$ measured by (3.23) does not change much, resulting in only small fluctuations in the q evolution trajectory.

The adaptive expected fitness gradient (3.20) with increasing q not only forces the trust region to move in the direction of the distribution that most likely represents areas with the highest packing density, but by doing so, it speeds up the rate of convergence. To illustrate this, we computed the L_1 norm of $\boldsymbol{\mu}_{\mathbf{F}_{1-\frac{1}{q_t}}} - \boldsymbol{\mu}$ at every optimization iteration for the values of $q = 6, \ldots, 600$. We scale the results such that the maximum at every iteration is not greater than one, precisely

$$(4.1) \qquad \| \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t, q) \|_{\mathrm{scaled}} = \frac{\| \boldsymbol{\mu}^t_{\mathbf{F}_{1-\frac{1}{q}}} - \boldsymbol{\mu}^t \|_1}{\| \boldsymbol{\mu}^t_{\mathbf{F}_{1-\frac{1}{q^{\max}}}} - \boldsymbol{\mu}^t \|_1},$$

where $q^{\max} = \mathrm{argmax}_q \| \boldsymbol{\mu}^t_{\mathbf{F}_{1-\frac{1}{q}}} - \boldsymbol{\mu}^t \|_1$, $q = 6, \ldots, 600$.

Although there are instances where $\| \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t, q_1) \|_{\mathrm{scaled}} \leq \| \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t, q_2) \|_{\mathrm{scaled}}$ for some $q_1 < q_2$ does not hold, Figure 4 shows that, on average, the higher the q, the larger the gradient size in $\boldsymbol{\theta}$-coordinates. This kind of behaviour is beneficial since the adaptive selection quantile accelerates convergences at later stages of the execution when an attractor has already been singled out.

The main principle of the algorithm is maximizing the expected fitness $J(\boldsymbol{\theta})$ (3.2). In our case, it means maximizing the packing density (2.5) or minimizing primitive cell volume. Figure 5 shows the evolution of the average density at each iteration defined by $\langle \rho \rangle = \frac{1}{N} \sum_{i=1}^{N} \frac{2\mathrm{AREA}(K)}{\mathbf{F}(\mathbf{x}_i)}$, where K is the regular octagon, N is the number of candidate solutions $\mathbf{x}_i$ sampled at each iteration and $\mathbf{F}(\mathbf{x})$ is the penalty function (D.2) with the objective function defined as the area of the unit cell $f(\mathbf{x}) := \mathrm{AREA}(\bar{\Lambda}(\mathbf{x}))$ and the constrain violation of the form $g(\mathbf{x}) := \mathrm{dist}(\mathcal{K}_{p2}(\mathbf{x}))$
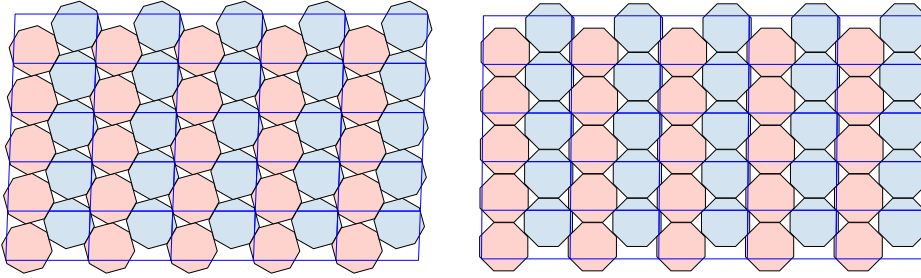
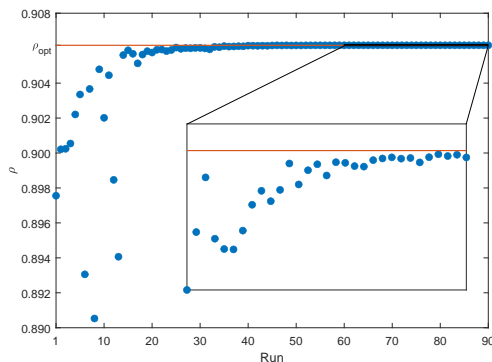Figure 9: (Left) Best and (Right) worst densest packing configurations found in 20 optimization runs.

Figure 10: Densities of the best solutions in each run of the packing refining process. ρ_{opt} denotes the theoretical optimal density.

defined by (D.7). The algorithm gradually increases the average density $\langle \rho \rangle$ with the maximum attained at the $4,378$th iteration.

Similar behaviour can be observed in the evolution of the maximum packing $\rho_{\max}$ found at each iteration, shown in Figure 5. The best solution was found at the $3,360$th iteration with packing density $\rho\left(\mathcal{K}_{p2}\right) = 0.897526117081202$ and minimal euclidean distance between octagons in the configuration $\mathrm{dist}\left(\mathcal{K}_{p2}\right) = 0.000385651690559$. A visualization of 25 cells from this configuration is presented in Figure 6. The difference from the theoretical optimal packing density ρ_{opt} defined as

$$(4.2) \qquad \Delta_{\mathcal{K}_{p2}} = \rho_{\mathrm{opt}} - \rho(\mathcal{K}_{p2}),$$

is $\Delta_{\mathcal{K}_{p2}} = 0.008637561562744$.

As the packing algorithm is stochastic in nature, it is unlikely to repeat the same result. Therefore, the maximum packing density attained during one execution can be considered a random variable by itself. To assess performance and robustness, we perform 20 consecutive runs with the same hyperparameters (Table 2) and the seed of the uniform distribution pseudorandom number generator used in the Gibbs sampler Algorithm B.1 initialized using system time. Figure 7 shows the evolutions of average density, and Figure 8 of maximum density. The Hodges-Lehmann estimator of the pseudomedian of the maximum packing density computed from the maximum densities attained in each of the 20 runs is $\hat{m} = 0.8970032$ with the 95% confidence interval based on Wilcoxon's signed rank test equal to CI $= (0.8959246, 0.8980686)$. On average, the maximum packing density was attained at the $4,624$th iteration. The
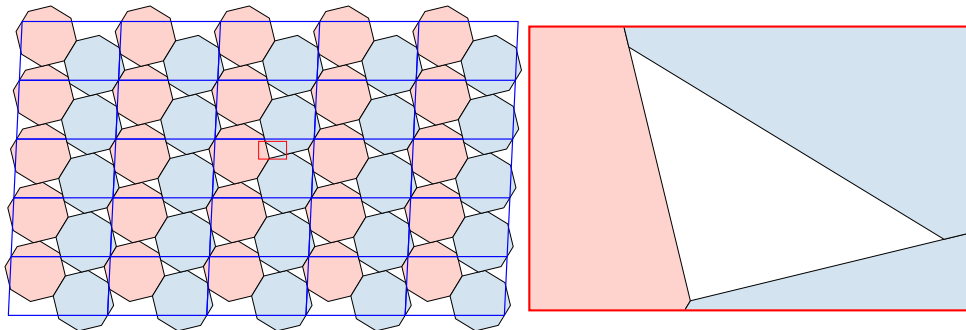
Figure 11: Refined solution. (Left) 25 cells from the whole configuration. (Right) The red rectangle marks an enlargement of the configuration segment.

highest maximal packing density configuration was attained in run 10 with packing density $\rho\left(\mathcal{K}_{p2}\right) = 0.899848387789551$, and the lowest maximal packing density was attained in run 20 with packing density $\rho\left(\mathcal{K}_{p2}\right) = 0.893366540122926$. Both solutions are shown in Figure 9. In a closer examination, it can be noticed that both configurations look similar, which is not surprising considering they both represent solutions from different global optima basins due to the multi-modality of the problem stemming from the symmetries of the regular octagon.

After the initial run, although the maximal packing density configuration found during the run of Figure 6 visually resembles the theoretical optimal packing of regular octagons, the difference of packing densities $\Delta_{\mathcal{K}_{p2}}$ (4.2) is in the 3rd decimal place. This can also be observed either by noticing the minimal euclidean distance between octagons in the packing dist $\left(\mathcal{K}_{p2}\right)$ defined by (D.7) or visually by enlarging a part of the packing, as is shown in Figure 6. Therefore, as introduced in subsection 3.5, we perform a refining process by taking the configuration with maximal packing density attained in the initial run (Figure 6) and creating an ϵ^r neighbourhood around this configuration's coordinates. In this way, we define a new configuration space with the boundaries given by (3.49) and rerun the algorithm with these new boundaries. At this point the optimization variables c_1, c_2 and ω_c lose their inherent periodicity and have to be treated as aperiodic in the boundary mapping. Figure 10 illustrates the convergence of the maximum density attained in each run during 90 runs of the decaying boundary neighbourhood. The highest packing density was attained at run 87 with the value $\rho\left(\mathcal{K}_{p2}\right) = 0.90616363432568$ and the theoretical optimum difference $\Delta_{\mathcal{K}_{p2}} = 3.1481\ 10^{-8}$. A visualization of this configuration is presented in Figure 11. For visual comparison, we include enlargement in Figure 11 of the same area of the output packing configuration as in the initial run (Figure 6).

5. Pentacene representation packings. We demonstrate how the densest CSG packings are intended to be used in a molecular CSP workflow on pentacene thin-films. First, a geometric representation of a molecule by a polytope and an invertible map from the molecule's atomic positions to the polytope's interior is constructed. Afterwards, the densest packings of this representation are obtained in various CSGs. Lastly, the parameters of output configurations are used as input parameters for CSP computations, reducing the CSP search only to the neighbourhood of these configurations.

Pentacene is a planar molecule consisting of five serially connected benzene rings [16], explored as an organic thin-film semiconductor [51]. Since the molecules in a
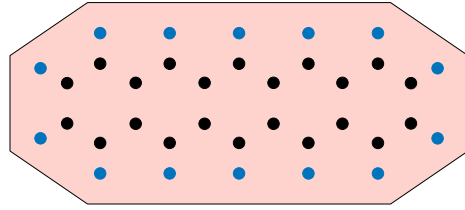
Figure 12: A geometric representation of pentacene. The dots symbolize atomic positions of (blue) hydrogen and (black) carbon.
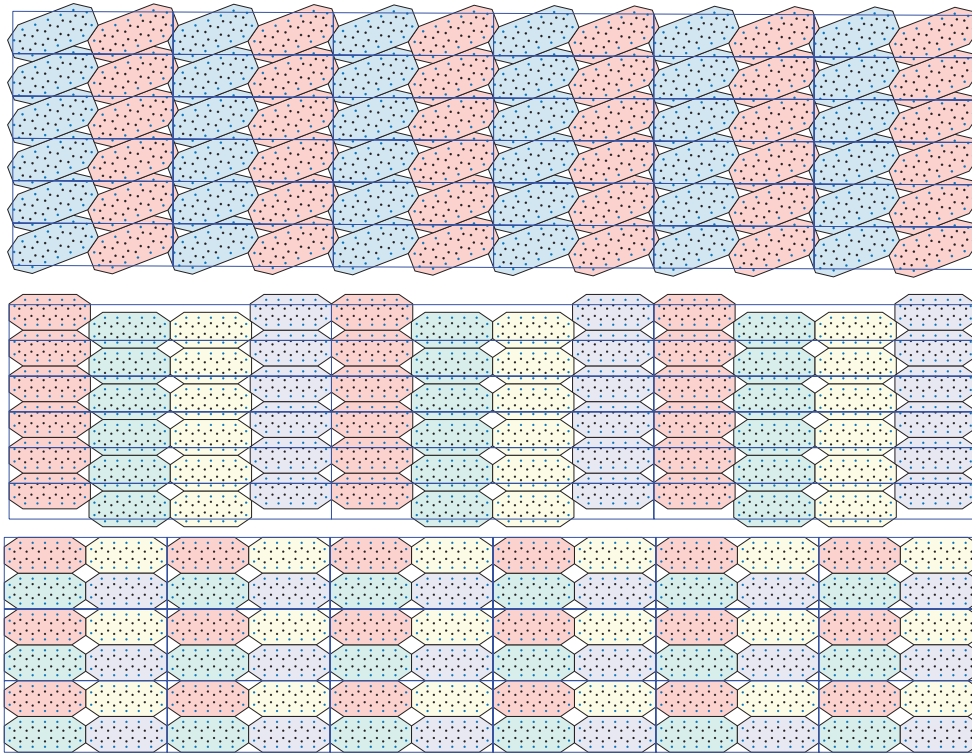

Figure 13: Visualization of the output configurations of the densest pentacene representation (top) $p2$-packing, (middle) cm-packing and (bottom) $p2mm$-packing. Colors represent CSG symmetry operations modulo lattice translations.

crystal do not touch due to repulsive intermolecular forces, we built a 2D representation of pentacene as the convex hull of the atomic positions of the molecule with an offset given by hydrogen's van der Waals radius of 1.09Å [65]. The result is an irregular octagon illustrated in Figure 12.

We employed the entropic trust region packing algorithm to search for this representation's densest plane group packings. Figure 13 presents output configurations of the densest $p2$, cm and $p2mm$ packings. The approximate density of the $p2$-packing is $\rho\left(\mathcal{K}_{p2}\right) = 0.953382110797399$ and resembles the configuration of single layer pentacene thin-film on graphite surface found in [18]. Moreover, this structure was found via simulation of self-assembly of a disordered system of pentacene molecules on graphene surface driven by the minimization of molecule-molecule in-

teractions using the Lennard–Jones potential [88]. The densities of the output configurations in the cm and $p2mm$ instances are $\rho\left(\mathcal{K}_{cm}\right) = 0.918715231405704$ and $\rho\left(\mathcal{K}_{p2mm}\right) = 0.910916580526171$. These configurations represent pentacene monolayer crystal phases on a Cu(110) surface found in [71].

6. Conclusions. The problem of molecular CSP is to predict stable periodic structures from the knowledge of the chemical composition of a molecule. The most straightforward formulation means finding minima on a complicated energy landscape induced by one of the many free energy potentials. This presents a formidable task for optimization methods and search algorithms, leading in many cases to an over–prediction of polymorphic structures [59]. Providing current CSP solvers with densely packed initial configurations in terms of the geometric representation of a molecule can significantly accelerate CSP, as opposed to random starting structures, especially due to the recent complete isometry invariants of periodic structures [3, 84, 83].

The densest packing of geometric shapes is a notoriously hard problem in discrete and computational geometry [76] and is used in a large body of work in solid-state physics modelling [75]. Since we are interested only in a particular class of periodic structures given by the crystallographic symmetry groups, in Section 2 we introduced a novel class of periodic packings, the CSG packings, by restricting possible packing configuration to CSG and formulated the densest CSG packing problem as a nonlinear bounded and constrained optimization problem.

Our motivation was to develop a search method for the densest packing for 2D and 3D CSP that is robust to a given geometric representation of a molecule. Moreover, the method needed to be agnostic to the search configuration space and the objective function properties. For example, in our experimental setting of octagon $p2$-packings, the non-overlap constraint incorporated into the penalty function is a continuous but not differentiable function, which renders the new objective function also not differentiable. In this manner, Section 3 restated the densest packing of polytopes via stochastic relaxation [33] and formulated a non-euclidean trust region method that solves this problem approximately. The resulting entropic trust region method performs updates along the geodesics on a statistical manifold where the trust region is given by KLD in a fashion similar to natural evolution strategies [85] and can be seen as an instance of the information geometry optimization framework [58].

The CSG restriction induces a toroidal topology on the packing configuration space. Therefore we perform the entropic trust region updates on a statistical manifold S consisting of a parametric family of probability distributions on an nD unit flat torus by extending the parameter space of multivariate von Mises distribution [50], introduced in Subsection 3.1. Moreover, the exponential family reparametrization of the extended multivariate von Mises distribution equips S with a dually flat structure [2].

Inspired by the simulated annealing control parameter [43], Subsection 3.2 introduced an adaptive quantile rewriting of the fitness into the entropic trust region update schedule to facilitate the search strategy. Consequently, the natural gradient of the adaptive selection quantile-based expected fitness points in the direction of the $(q-1)$th q quantile of the fitness transformed random vector, serving as an adaptive step length method.

The natural gradient descent [1] and the generalized proximal minimization algorithm [17] share a common ground due to the Bregman divergence characterized by the exponential family log-partition function discussed in Subsection 3.3. In Subsection 3.4 we used this knowledge together with the dual structure given by the

	$\rho\left(\mathcal{K}_G\right)$	dist $\left(\mathcal{K}_G\right)$	Δ
regular octagon in $p2$	0.90616363432568	3.1481 10^{-8}	4.4318 10^{-8}
regular pentagon in pg	0.92131060131385	2.1811 10^{-8}	7.2852 10^{-8}
regular heptagon in $p2gg$	0.89269066997639	8.2820 10^{-10}	1.6150 10^{-8}
irregular pentagon in $p4$	0.99999999503997	1.5809 10^{-9}	4.9600 10^{-9}
regular hexagon in $p3$	0.99999993380570	1.9423 10^{-9}	6.6194 10^{-8}
$30-60-90$ triangle in $p6mm$	0.99999999871467	4.4161 10^{-11}	1.2853 10^{-9}

Table 1: Packing density $\rho\left(\mathcal{K}_G\right)$, the minimal distance between a pair of polygons in a packing dist $\left(\mathcal{K}_G\right)$ and packing density difference from the theoretical optimum Δ for the presented test cases.

exponentially reparametrized extended multivariate von Mises statistical model and examined the geometry of the adaptive selection quantile equipped trust region. Embedding the statistical model S to a statistical model consisting of probability distributions derived from S by truncating S at the $(q-1)$th q-quantile of the fitness, where for every fixed q, S becomes a submanifold of codimension 1, we show that the resulting dual geodesic flow induced by the trust region search directions performs minimax of KLD between two hypersurfaces, one given by the statistical model S and the other by S truncated at the $(q-1)$th q-quantile of the fitness. Moreover, this minimax maximizes the stochastic dependence between the elements of the extended multivariate von Mises distributed random vector, measured by multi-information [72] or total correlation [82], providing the entropic trust region with even greater model interpretability.

Applying the proposed algorithm to the densest $p2$-packing of regular octagons, presented in Section 4, showed competitive performance, even considering the relatively low number of samples used in the Monte–Carlo estimates, compared to the order of the exponential family rewriting of the extended multivariate von Mises distribution or dimensionality of the statistical model S, and the multi-modality of the optimization landscape. Furthermore, through the refining solution process, the algorithm achieved high accuracy measured by differences to the known theoretical optima. Moreover, the output configuration Figure 11 shows higher symmetry of the densest regular octagon packing than that of a lattice packing. The algorithm performed equally well when applied to the densest packings of regular and irregular convex polygons for which theoretical optimal solutions are known Table 1. In all cases, the difference from the theoretical optimal solutions is lower than 10^{-7} and potentially could yield better approximations provided the refining process is allowed to run longer.

Although we chose plane group packings for demonstrating the behavior and performance of the entropic trust region, the optimization algorithm is constructed to search CSGs of arbitrary dimensions. For example, in the setting of the densest space group packing of a convex polyhedron for the triclinic crystal system, the configuration space constitutes a 12D torus given by three fractional coordinates of the polyhedron centroid, three angles of rotation of the polyhedron around the respective axes, three lengths of primitive cell edges and three angles between primitive cell edges. Moreover, the algorithm is implemented modularly, with objective function and constraints user-specifiable as inputs which render the algorithm applicable to any nonlinear bounded constrained optimization.

However, there are a few caveats regarding higher dimensional packing. The main computational bottleneck is the extended multivariate von Mises distribution

Gibbs sampler (Appendix B.3) which needs to scale better. By increasing the CSG dimension, the dimensionality of the configuration space rises polynomially, and more efficient sampling methods are necessary. Since the standard multivariate von Mises model is a stationary distribution of a Langevin diffusion stochastic differential equation [30], a natural approach is to construct a Metropolis-adjusted Langevin algorithm [34]. Additionally, the stabilization of the Fisher metric tensor estimate is another caveat related to the increased dimensionality of the higher dimensional CSG packing problem. In the current implementation, a spectral radius scaling of the Fisher matrix (Appendix D.5) is used to improve the stability of the dynamical system underlying the Entropic trust region packing algorithm. A strategy to further improve the stability and reduce the number of samples necessary for accurate estimation is to regard the scaled Fisher metric tensor as a diffusion matrix [27] and derive conditions when the resulting Riemannian gradient defines a contraction mapping. The second most significant computational bottleneck is the overlap constraint violation computation (Appendix D.3) for a given CSG configuration. Here efficiency can be likewise improved by various heuristics. For instance, for two polytopes in a CSG configuration, it is not necessary to compute the degree of overlap if they do not intersect. Further, if the polytope circumspheres do not intersect, the polytopes do not intersect. Since the computation of sphere overlap is just one operation, the overlap constraint violation can be significantly improved for cases where polytope circumspheres do not intersect, compared to the full search for the separating hyperplane.

Subsequent work is to incorporate the presented search method into existing CSP solvers to guide CSP tasks. This requires assigning a geometric representation to a molecule that can be done either manually by examining intrinsic topological properties given by the chemical composition of a molecule [66, 80] or automatically by, for example, taking the convex hull of the point set generated by the atomic coordinates of each atom as we demonstrated in Section 5. The situation is more complicated in the 2D CSP case since the molecule is usually defined as being embedded in 3D Euclidean space. Thus to construct a polygon representation of the molecule, one needs to choose a projection onto the 2D Euclidean space.

REFERENCES

[1] S. AMARI, *Natural gradient works efficiently in learning*, Neural Computation, (1998), pp. 251–276.
[2] S. AMARI AND H. NAGAOKA, *Methods of Information Geometry.*, American Mathematical Society., 2nd ed., 2000.
[3] O. ANOSOVA AND V. KURLIN, *An isometry classification of periodic point sets*, in Lecture Notes in Computer Science (Proceedings of DGMM), vol. 12708, 2021, pp. 229–241.
[4] S. ATKINSON, Y. JIAO, AND S. TORQUATO, *Maximally dense packings of two-dimensional convex and concave noncircular particles*, Physical Review E, 86 (2012), p. 031302.
[5] N. AY, *An information–geometric approach to a theory of pragmatic structuring*, The Annals of Probability, 30 (2002), pp. 416–436.
[6] N. AY AND A. KNAUF, *Maximizing multi-information*, Kybernetika, 42 (2006), pp. 517–538.
[7] A. BANERJEE, S. MERUGU, I. S. DHILLON, J. GHOSH, AND J. LAFFERTY, *Clustering with bregman divergences.*, Journal of machine learning research, 6 (2005).
[8] O. BARNDORFF-NIELSEN, *Information and Exponential Families: In Statistical Theory*, John Wiley & Sons, 2014.

[9] A. Beck and M. Teboulle, *Mirror descent and nonlinear projected subgradient methods for convex optimization*, Operations Research Letters, (2003), pp. 167–175.

[10] U. Betke and M. Henk, *Densest lattice packings of 3-polytopes*, Computational Geometry, 16 (2000), pp. 157–186.

[11] H.-G. Beyer, *The Theory of Evolution Strategies*, Springer, 2001.

[12] H.-G. Beyer, *Convergence analysis of evolutionary algorithms that are based on the paradigm of information geometry*, Evolutionary Computation, (2014).

[13] A. Bezdek, *A remark on the packing density in the 3-space*, Intuitive geometry, 63 (1994), p. 17.

[14] L. M. Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR computational mathematics and mathematical physics, 7 (1967), pp. 200–217.

[15] C. P. Brock, T. Hahn, H. Wondratschek, U. Müller, U. Shmueli, E. Prince, A. Authier, V. Kopský, D. Litvin, E. Arnold, et al., *International tables for crystallography volume a: Space-group symmetry*, 2016.

[16] R. Campbell, J. M. Robertson, and J. Trotter, *The crystal and molecular structure of pentacene*, Acta crystallographica, 14 (1961), pp. 705–711.

[17] Y. Censor and S. A. Zenios, *Proximal minimization algorithm with d-functions*, Journal of Optimization Theory and Applications, 73 (1992), pp. 451–464.

[18] W. Chen, H. Huang, A. Thye, and S. Wee, *Molecular orientation transition of organic thin films on graphite: the effect of intermolecular electrostatic and interfacial dispersion forces*, Chemical communications, (2008), pp. 4276–4278.

[19] N. Chernov, Y. Stoyan, and T. Romanova, *Mathematical model and efficient algorithms for object packing problem*, Computational Geometry, (2010).

[20] C. A. C. Coello, G. B. Lamont, D. A. Van Veldhuizen, et al., *Evolutionary algorithms for solving multi-objective problems*, vol. 5, Springer, 2007.

[21] R. G. Cowell, P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter, *Probabilistic networks and expert systems: Exact computational methods for Bayesian networks*, Springer Science & Business Media, 2007.

[22] D. Cui, M. Ebrahimi, J. M. Macleod, and F. Rosei, *Template-driven dense packing of pentagonal molecules in monolayer films*, Nano Letters, 18 (2018), pp. 7570–7575.

[23] S. Das, A. Sebastian, E. Pop, C. J. McClellan, A. D. Franklin, T. Grasser, T. Knobloch, Y. Illarionov, A. V. Penumatcha, J. Appenzeller, et al., *Transistors based on two-dimensional materials for future integrated circuits*, Nature Electronics, 4 (2021), pp. 786–799.

[24] J. de Graaf, R. van Roij, and M. Dijkstra, *Dense regular packings of irregular nonconvex particles*, Physical Review Letters, 107 (2011), p. 155501.

[25] K. Deb, *An efficient constraint handling method for genetic algorithms*, Computer methods in applied mechanics and engineering, 186 (2000), pp. 311–338.

[26] D. Ecija, S. Vijayaraghavan, W. Auwärter, S. Joshi, K. Seufert, C. Aurisicchio, D. Bonifazi, and J. V. Barth, *Two-dimensional short-range disordered crystalline networks from flexible molecular modules*, ACS nano, 6 (2012), pp. 4258–4265.

[27] R. Fioresi, P. Chaudhari, and S. Soatto, *A geometric interpretation of stochastic gradient descent using diffusion metrics*, Entropy, 22 (2020), p. 101.

[28] L. Flatley and F. Theil, *Face-centered cubic crystallization of atomistic configurations*, Archive for Rational Mechanics and Analysis, 218 (2015), pp. 363–416.

[29] J. D. Gale and A. L. Rohl, *The general utility lattice program (gulp)*, Molecular Simulation, 29 (2003), pp. 291–341.

[30] E. García-Portugués, M. Sørensen, K. V. Mardia, and T. Hamelryck, *Langevin diffusions on the torus: estimation and applications*, Statistics and Computing, 29 (2019), pp. 1–22.

[31] R. Gatto, *Some computational aspects of the generalized von mises distribution*, Statistics and computing, (2008).

[32] R. Gatto and S. R. Jammalamadaka, *The generalized von mises distribution*, Statistical Methodology, (2007).

[33] S. Geman and D. Geman, *Stochastic relaxation, gibbs distributions, and the bayesian restoration of images*, IEEE Transactions on pattern analysis and machine intelligence, (1984), pp. 721–741.

[34] M. Girolami and B. Calderhead, *Riemann manifold langevin and hamiltonian monte carlo methods*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73 (2011), pp. 123–214.

[35] J. Hafner, *Ab-initio simulations of materials using vasp: Density-functional theory and be-

yond, Journal of computational chemistry, 29 (2008), pp. 2044–2078.

[36] T. Hales and W. Kusner, *Packings of regular pentagons in the plane*, arXiv:1602.07220, (2016).

[37] T. C. Hales, *A proof of the kepler conjecture*, Annals of mathematics, 162 (2005), pp. 1065–1185.

[38] N. Hansen and A. Ostermeier, *Completely derandomized self-adaptation in evolution strategies*, Evolutionary computation, 9 (2001), pp. 159–195.

[39] E. Huber and W. Barth, *Surface-to-surface intersection with complete and guaranteed results*, Springer Science & Business Media, 2013.

[40] S. R. Jammalamadaka and A. Sengupta, *Topics in circular statistics*, vol. 5, world scientific, 2001.

[41] P. E. Jupp and K. V. Mardia, *A general correlation coefficient for directional data and related regression problems*, Biometrika, 67 (1980), pp. 163–173.

[42] G. Kuperberg and W. Kuperberg, *Double-lattice packings of convex bodies in the plane, computational geometry*, Discrete & Computational Geometry, (1990).

[43] P. J. M. v. Laarhoven and E. H. L. Aarts, *Simulated Annealing: Theory and Application*, Springer, 1987.

[44] S. L. Lauritzen, *Statistical manifolds*, Differential geometry in statistical inference, 10 (1987), pp. 163–216.

[45] J. M. Lee, *Distributions and foliations*, in Introduction to Smooth Manifolds, Springer, 2013, pp. 490–514.

[46] M. Lewin and X. Blanc, *The crystallization conjecture: a review*, EMS Surveys in Mathematical Sciences, 2 (2015), pp. 255–306.

[47] R. Linsker, *A local learning rule that enables information maximization for arbitrary input distributions*, Neural Computation, 9 (1997), pp. 1661–1665.

[48] L. Malagò, M. Matteucci, and G. Pistone, *Towards the geometry of estimation of distribution algorithms based on the exponential family*, Proceedings of the 11th workshop proceedings on Foundations of genetic algorithms, (2011), pp. 230–242.

[49] K. V. Mardia, *Statistics of directional data*, J. R. Statist. Soc., (1975).

[50] K. V. Mardia, G. Hughes, C. C. Taylor, and H. Singh, *A multivariate von mises distribution with applications to bioinformatics*, Canadian Journal of Statistics, (2008).

[51] J. Mei, Y. Diao, A. L. Appleton, L. Fang, and Z. Bao, *Integrated materials design of organic semiconductors for field-effect transistors*, Journal of the American Chemical Society, 135 (2013), pp. 6724–6746.

[52] W. J. Miller, *Symmetry groups and their applications*, Accademic press, 1972.

[53] D. M. Mount, *The densest double-lattice packing of a convex polygon*, Discrete and Computational Geometry: Papers from the DIMACS Special Year, 6 (1991), pp. 245–262.

[54] D. M. Mount and R. Silverman, *Packing and covering the plane with translates of a convex polygon*, Journal of Algorithms, 11 (1990), pp. 564–580.

[55] T. K. Mukhopadhyay and A. Datta, *Ordering and dynamics for the formation of two-dimensional molecular crystals on black phosphorene*, The Journal of Physical Chemistry C, 121 (2017), pp. 10210–10223.

[56] A. S. Nemirovskij and D. B. Yudin, *Problem complexity and method efficiency in optimization*, (1983).

[57] J. Nocedal and J. W. Wright, *Numerical Optimization*, Springer, 2nd ed., 2000.

[58] Y. Ollivier, L. Arnold, A. Auger, and N. Hansen, *Information-geometric optimization algorithms: A unifying picture via invariance principles*, Journal of Machine Learning Research, (2017), pp. 251–276.

[59] S. L. Price, *Is zeroth order crystal structure prediction (csp_0) coming to maturity? what should we aim for in an ideal crystal structure prediction code?*, Faraday discussions, 211 (2018), pp. 9–30.

[60] G. Raskutti and S. Mukherjee, *The information geometry of mirror descent*, IEEE Transactions on Information Theory, (2015), pp. 1451 – 1457.

[61] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*, Springer, 2nd ed., 2004.

[62] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*, vol. 317, Springer Science & Business Media, 2009.

[63] C. A. Rogers, *The closest packing of convex two-dimensional domains*, Acta Math., 86 (1951), pp. 309–321.

[64] C. A. Rogers, *Packing and Covering*, Cambridge University Press, 1964.

[65] R. S. Rowland and R. Taylor, *Intermolecular nonbonded contact distances in organic crystal structures: Comparison with distances expected from van der waals radii*, The Journal of Physical Chemistry, 100 (1996), pp. 7384–7391.

[66] V. Santolini, M. Miklitz, E. Berardo, and K. E. Jelfs, *Topological landscapes of porous organic cages*, Nanoscale, 9 (2017), pp. 5280–5298.

[67] T. Schilling, S. Pronk, B. Mulder, and D. Frenkel, *Monte carlo study of hard pentagons*, Physical Review E, 71 (2005), p. 036138.

[68] J. E. Shore and W. J. Rodney, *Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy*, IEEE Transactions on Information theory, (1980).

[69] F. M. Silva and L. B. Almeida, *Acceleration techniques for the backpropagation algorithm*, Neural Networks, (1990).

[70] H. Singh, V. Hnizdo, and E. Demchuk, *Probabilistic model for two dependent circular variables*, Biometrika, 89 (2002), pp. 719–723.

[71] S. Söhnchen, S. Lukas, and G. Witte, *Epitaxial growth of pentacene films on cu (110)*, The Journal of chemical physics, 121 (2004), pp. 525–534.

[72] M. Studený and J. Vejnarová, *The multiinformation function as a tool for measuring stochastic dependence*, Learning in Graphical Models, (1998).

[73] M. Teboulle, *Entropic proximal mappings with applications to nonlinear programming*, Mathematics of operations research, (1992).

[74] F. Theil, *A proof of crystallization in two dimensions*, Communications in Mathematical Physics, 262 (2006), pp. 209–236.

[75] S. Torquato, *Perspective: Basic understanding of condensed phases of matter via packing models*, The Journal of chemical physics, 149 (2018), p. 020901.

[76] C. D. Toth, J. O'Rourke, and J. E. Goodman, *Handbook of discrete and computational geometry*, CRC press, 2017.

[77] L. F. Tóth, *Lagerungen in der Ebene auf der Kugel und im Raum*, vol. 65, Springer-Verlag, 2013.

[78] H. L. V. Trees and K. L. Bell, *Bayesian bounds for parameter estimation and nonlinear filtering/tracking*, Wiley-IEEE press New York, 2007.

[79] D. J. Wales, *Exploring energy landscapes*, Annual review of physical chemistry, 69 (2018), pp. 401–425.

[80] H. Wang, L.-P. Zhou, Y. Zheng, K. Wang, B. Song, X. Yan, L. Wojtas, X.-Q. Wang, X. Jiang, M. Wang, et al., *Double-layered supramolecular prisms self-assembled by geometrically non-equivalent tetratopic subunits*, Angewandte Chemie, 133 (2021), pp. 1318–1325.

[81] H.-W. Wang and S.-f. Sui, *Pentameric two-dimensional crystallization of rabbit c-reactive protein on lipid monolayers*, Journal of structural biology, 127 (1999), pp. 283–286.

[82] S. Watanabe, *Information theoretical analysis of multivariate correlation*, IBM Journal of Research and Development, 4 (1960), pp. 66–82.

[83] D. Widdowson and V. Kurlin, *Resolving the data ambiguity for periodic crystals*, Advances Neural Inform. Proc. Systems (NeurIPS), 35 (2022).

[84] D. Widdowson, M. Mosca, A. Pulido, A. Cooper, and V. Kurlin, *Average minimum distances of periodic point sets - fundamental invariants for mapping all periodic crystals*, MATCH Comm. Math. Comp. Chemistry, 87 (2022), pp. 529–559.

[85] D. Wierstra, T. Schaul, T. Glasmachers, Y. Sun, J. Peters, and J. Schmidhuber, *Natural evolution strategies*, Journal of Machine Learning Research, (2014), pp. 949–980.

[86] S. Willard, *General topology*, Addison-Wesley, 1970.

[87] S. M. Woodley, G. M. Day, and R. Catlow, *Structure prediction of crystals, surfaces and nanoparticles*, Philosophical Transactions of the Royal Society A, 378 (2020), p. 20190600.

[88] Y. Zhao, Q. Wu, Q. Chen, and J. Wang, *Molecular self-assembly on two-dimensional atomic crystals: insights from molecular dynamics simulations*, The journal of physical chemistry letters, 6 (2015), pp. 4518–4524.

Appendix A. Estimating natural gradients. Generally, the explicit computation of integrals for the natural gradient $\widetilde{\nabla}J(\boldsymbol{\theta})$ in (3.9) is not possible. For example, in our case the normalizer $Z(\boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{D})$ in (3.10) is unknown. The standard workaround in these situations is to use Monte–Carlo methods.

For exponential families, the situation becomes even simpler. The Fisher information matrix $\mathcal{I}$ (3.5) equals the variance of sufficient statistic $\mathbf{t}$ and then the estimate takes the following form

$$(A.1) \qquad \widehat{\mathcal{I}_{\boldsymbol{\theta}}} = \widehat{\mathrm{VAR}}\left[\mathbf{t}(\mathbf{x})|\boldsymbol{\theta}\right],$$

where $\widehat{\mathrm{VAR}}$ denotes the sample covariance matrix.

The $\boldsymbol{\theta}$ gradient of $J(\boldsymbol{\theta})$ in the case of exponential families is obtained by differentiating (3.2) as

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \int_{\mathbf{X}} \mathbf{F}(\mathbf{x})\left[\mathbf{t}(\mathbf{x}) - \nabla_{\boldsymbol{\theta}}\psi\left(\boldsymbol{\theta}\right)\right] dP(\boldsymbol{\theta}),$$

which is the expected value of $\mathbf{F}(\mathbf{x})\left[\mathbf{t}(\mathbf{x}) - \nabla_{\boldsymbol{\theta}}\psi\left(\boldsymbol{\theta}\right)\right]$ with respect to the probability distribution $P(\boldsymbol{\theta})$. The standard gradient of the free energy $\psi\left(\boldsymbol{\theta}\right)$ equals the expected value of the sufficient statistic $\nabla_{\boldsymbol{\theta}}\psi\left(\boldsymbol{\theta}\right) = \int_{\mathbf{X}} \mathbf{t}(\mathbf{x})dP(\boldsymbol{\theta})$ and the estimate of the standard gradient of the expected fitness can then take the form

$$(A.2) \qquad \widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})} = \widehat{E}\left[\mathbf{F}(\mathbf{x})\left(\mathbf{t}(\mathbf{x}) - \widehat{E}\left[\mathbf{t}(\mathbf{x})\right]\right) | \boldsymbol{\theta}\right],$$

where $\widehat{E}$ denotes the sample mean. The expression (A.2) can be simplified further and receive a clear geometric interpretation using the selection quantile introduced in subsection 3.2.

By combining (A.1) and (A.2), the natural gradient estimate takes the form

$$(A.3) \qquad \widehat{\widetilde{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})}} = \widehat{\mathcal{I}_{\boldsymbol{\theta}}}^{-1} \widehat{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})}.$$

The above expression is also used to estimate the natural gradient for exponential families in [48] and [58].

Appendix B. Toroidal distributions.
The general bivariate von Mises distribution density function has the following form [49]

$$(B.1) \quad f(\theta_1, \theta_2|\kappa_1, \kappa_2, \mu_1, \mu_2, \mathbf{A}) = \frac{1}{Z(\kappa_1, \kappa_2, \mu_1, \mu_2, \mathbf{A})} \exp\left\{\kappa_1 \cos(\theta_1 - \mu_1) + \right.$$

$$\left. +\kappa_2 \cos(\theta_2 - \mu_2) + \begin{bmatrix}\cos(\theta_1 - \mu_1)\\ \sin(\theta_1 - \mu_1)\end{bmatrix}^{\mathrm{T}} \mathbf{A} \begin{bmatrix}\cos(\theta_2 - \mu_2)\\ \sin(\theta_2 - \mu_2)\end{bmatrix}\right\},$$

where $0 \le \theta_1, \theta_2 < 2\pi$ represent angles of corresponding unit circles S^1 of the product space $T^2 = S^1 \times S^1$, $Z(\kappa_1, \kappa_2, \mu_1, \mu_2, \mathbf{A})$ is the normalizer, μ_1, μ_2 are the mean direction parameters, κ_1, κ_2 are concentration parameters and $\mathbf{A}$ is a 2×2 matrix representing dependencies between angles θ_1, θ_2.

Based on the sine submodel of the general bivariate von Mises model (B.1), [50] defined a probability distribution on an nD torus T^n with probability density

$$(B.2) \qquad f(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda}) = \frac{1}{Z(\boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda})} \exp\left\{\boldsymbol{\kappa}^{\mathrm{T}} c(\boldsymbol{\theta} - \boldsymbol{\mu}) + \frac{1}{2} s(\boldsymbol{\theta} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Lambda} s(\boldsymbol{\theta} - \boldsymbol{\mu})\right\}$$

where

$$c(\boldsymbol{\theta} - \boldsymbol{\mu}) = [cos(\theta_1 - \mu_1), \ldots, cos(\theta_n - \mu_n)]^{\mathrm{T}},$$
$$s(\boldsymbol{\theta} - \boldsymbol{\mu}) = [sin(\theta_1 - \mu_1), \ldots, sin(\theta_n - \mu_n)]^{\mathrm{T}},$$
$$-\pi \le \theta_i \le \pi, \ -\pi \le \mu_i \le \pi, \ 0 \le \kappa_i, \ -\infty \le \lambda_{ij} \le \infty,$$
$$\boldsymbol{\Lambda}_{ij} = \lambda_{ij} = \lambda_{ji}, \ i \ne j, \ , \lambda ii = 0.$$

for $i, j = 1, \ldots, n$.

Following the full bivariate von Mises model (B.1), we extend the multivariate von Mises model (B.2) to the full model with probability density

$$(\text{B.3}) \quad f(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{D}) = \frac{1}{Z(\boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{D})} \exp\left\{ \boldsymbol{\kappa}^{\mathrm{T}} c(\boldsymbol{\theta} - \boldsymbol{\mu}) + \begin{bmatrix} c(\boldsymbol{\theta} - \boldsymbol{\mu}) \\ s(\boldsymbol{\theta} - \boldsymbol{\mu}) \end{bmatrix}^{\mathrm{T}} \mathbf{D} \begin{bmatrix} c(\boldsymbol{\theta} - \boldsymbol{\mu}) \\ s(\boldsymbol{\theta} - \boldsymbol{\mu}) \end{bmatrix} \right\}$$

where $\mathbf{D}$ is $2n \times 2n$ matrix with no restrictions whatsoever.

Although there are no restrictions on $\mathbf{A}$ in (B.1) only a few specific submodels are considered in actual applications [40] due to the difficulty in the statistical interpretation of the distribution parameters and a more direct relationship with the bivariate normal distribution. The same applies to the multivariate von Mises distribution (B.2) where the matrix $\text{diag}(\boldsymbol{\kappa}) - \boldsymbol{\Lambda}$ can be interpreted as the precision matrix of a multivariate normal distribution. However, our application of the full multivariate von Mises model (B.3) is not dependent upon statistical interpretations of the parameters but is related to the interpretation of the probabilistic trust region method (3.1). The multivariate von Mises model (B.2) is a submodel of the full multivariate model (B.3), and the additional degrees of freedom enable the distribution to better approximate the optimization landscape induced by the expected fitness (3.2).

B.1. Exponential reformulation of the full multivariate von Mises distribution. By restricting $\boldsymbol{\kappa} > \mathbf{0}$ and using trigonometric identities, we rewrite the full multivariate von Mises model (B.3) to the exponential family parameterization

$$(\text{B.4}) \quad f(\boldsymbol{\theta}|\boldsymbol{\eta}, \mathbf{E}) = \exp\left\{ \begin{bmatrix} c(\boldsymbol{\theta}) \\ s(\boldsymbol{\theta}) \end{bmatrix}^{\mathrm{T}} \boldsymbol{\eta} + \text{vec}\left(\begin{bmatrix} c(\boldsymbol{\theta}) \\ s(\boldsymbol{\theta}) \end{bmatrix} \begin{bmatrix} c(\boldsymbol{\theta}) \\ s(\boldsymbol{\theta}) \end{bmatrix}^{\mathrm{T}} \right)^{\mathrm{T}} \text{vec}(\mathbf{E}) - \psi(\boldsymbol{\eta}, \mathbf{E}) \right\}$$

where $\text{vec}(\cdot)$ denotes vectorization operation, with the canonical exponential family parameters $(\boldsymbol{\eta}, \mathbf{E})$ given by

$$(\text{B.5a}) \qquad\qquad \boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\kappa} \odot c(\boldsymbol{\mu}) \\ \boldsymbol{\kappa} \odot s(\boldsymbol{\mu}) \end{bmatrix},$$

$$(\text{B.5b}) \qquad\qquad \mathbf{E} = \begin{bmatrix} \mathbf{E}^{cc} & \mathbf{E}^{cs} \\ \mathbf{E}^{sc} & \mathbf{E}^{ss} \end{bmatrix},$$

$$\mathbf{E}^{cc} = \mathbf{D}^{cc} \odot c(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathrm{T}} - \mathbf{D}^{cs} \odot c(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathrm{T}} - \mathbf{D}^{sc} \odot s(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathrm{T}} + \mathbf{D}^{ss} \odot s(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathrm{T}},$$
$$\mathbf{E}^{cs} = \mathbf{D}^{cc} \odot c(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathrm{T}} + \mathbf{D}^{cs} \odot c(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathrm{T}} - \mathbf{D}^{sc} \odot s(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathrm{T}} - \mathbf{D}^{ss} \odot s(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathrm{T}},$$
$$\mathbf{E}^{sc} = \mathbf{D}^{cc} \odot s(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathrm{T}} - \mathbf{D}^{cs} \odot s(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathrm{T}} + \mathbf{D}^{sc} \odot c(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathrm{T}} - \mathbf{D}^{ss} \odot c(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathrm{T}},$$
$$\mathbf{E}^{ss} = \mathbf{D}^{cc} \odot s(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathrm{T}} + \mathbf{D}^{cs} \odot s(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathrm{T}} + \mathbf{D}^{sc} \odot c(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathrm{T}} + \mathbf{D}^{ss} \odot c(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathrm{T}},$$

where $\odot$ denotes Hadamard product and

(B.6)
$$\mathbf{D} = \begin{bmatrix} \mathbf{D}^{cc} & \mathbf{D}^{cs} \\ \mathbf{D}^{sc} & \mathbf{D}^{ss} \end{bmatrix}$$

is the interaction matrix in the full multivariate von Mises model (B.3).

Clearly, the exponential rewriting of the full multivariate von Mises model is over-parametrized. In order to guarantee that the Fisher information $\mathcal{I}$ (3.5) is positive definite, the canonical exponential family parameters of the transformed concentration, circular mean and interaction parameters need to be affinely independent since the Fisher information matrix is a Gram matrix of log-likelihood differentials of (B.5) with respect to the inner product given by (B.5). Thus, we can reparametrize the model (B.4) to the minimal canonical form using the observation that

(B.7)
$$e_{ij}^{cc} \cos(\theta_i) \cos(\theta_j) + e_{ji}^{cc} \cos(\theta_j) \cos(\theta_i) = (e_{ij}^{cc} + e_{ji}^{cc}) \cos(\theta_i) \cos(\theta_j),$$
$$e_{ij}^{sc} \sin(\theta_i) \cos(\theta_j) + e_{ji}^{cs} \cos(\theta_j) \sin(\theta_i) = (e_{ij}^{sc} + d_{ji}^{cs}) \sin(\theta_i) \cos(\theta_j),$$
$$e_{ij}^{ss} \sin(\theta_i) \sin(\theta_j) + e_{ji}^{cc} \sin(\theta_j) \sin(\theta_i) = (e_{ij}^{ss} + e_{ji}^{ss}) \sin(\theta_i) \sin(\theta_j),$$
$$e_{ii}^{cc} (\cos(\theta_i))^2 + e_{ii}^{ss} (\sin(\theta_i))^2 = e_{ii}^{ss} + (e_{ii}^{cc} - e_{ii}^{ss})(\cos(\theta_i))^2,$$

for $i, j = 1, \ldots, n$, where e_{ij}^{cc}, e_{ij}^{ss}, e_{ij}^{cs}, e_{ij}^{sc} are elements of $\mathbf{E}^{cc}$, $\mathbf{E}^{ss}$, $\mathbf{E}^{cs}$, $\mathbf{E}^{sc}$ submatrices of $\mathbf{E}$ in (B.5b). Based on the introduced reparametrization (B.7), $\mathbf{E}$ becomes symmetric and $e_{ii}^{cc} = -e_{ii}^{ss}$.

As due to the symmetry of $\mathbf{E}$ the non–diagonal parameters are counted twice, we remove this redundancy by scaling $\mathbf{D}$ in (B.3) by the factor of 2. Now the transformation (B.5b) becomes

$$\mathbf{E}^{cc} = \tfrac{1}{2}\mathbf{D}^{cc} \odot c(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathsf{T}} - \tfrac{1}{2}\mathbf{D}^{cs} \odot c(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathsf{T}} + \tfrac{1}{2}\mathbf{D}^{ss} \odot s(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathsf{T}} - \tfrac{1}{2}\,\mathbf{D}^{cs\,\mathsf{T}} \odot s(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathsf{T}}$$

$$\mathbf{E}^{cs} = \tfrac{1}{2}\mathbf{D}^{cs} \odot c(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathsf{T}} + \tfrac{1}{2}\mathbf{D}^{cc} \odot c(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathsf{T}} - \tfrac{1}{2}\mathbf{D}^{ss} \odot s(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathsf{T}} - \tfrac{1}{2}\,\mathbf{D}^{cs\,\mathsf{T}} \odot s(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathsf{T}}$$

$$\mathbf{E}^{ss} = \tfrac{1}{2}\mathbf{D}^{ss} \odot c(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathsf{T}} + \tfrac{1}{2}\mathbf{D}^{cc} \odot s(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathsf{T}} + \tfrac{1}{2}\mathbf{D}^{cs} \odot s(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathsf{T}} + \tfrac{1}{2}\,\mathbf{D}^{cs\,\mathsf{T}} \odot c(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathsf{T}}$$

and as a consequence, the full multivariate von Mises model (B.3) now becomes
(B.9)
$$f(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{D}) = \frac{1}{Z(\boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{D})} \exp\left\{ \boldsymbol{\kappa}^{\mathsf{T}} c(\boldsymbol{\theta} - \boldsymbol{\mu}) + \frac{1}{2} \begin{bmatrix} c(\boldsymbol{\theta} - \boldsymbol{\mu}) \\ s(\boldsymbol{\theta} - \boldsymbol{\mu}) \end{bmatrix}^{\mathsf{T}} \mathbf{D} \begin{bmatrix} c(\boldsymbol{\theta} - \boldsymbol{\mu}) \\ s(\boldsymbol{\theta} - \boldsymbol{\mu}) \end{bmatrix} \right\},$$

where $\mathbf{D}$ is $2n \times 2n$ matrix with the same structure as $\mathbf{E}$ in (B.5b), that is $\mathbf{D}$ is symmetric with the diagonal elements d_{ii}^{cc}, d_{ii}^{ss} of the $\mathbf{D}^{cc}$, $\mathbf{D}^{ss}$ submatrices of $\mathbf{D}$ in (B.6) related by $d_{ii}^{cc} = -d_{ii}^{ss}$.

The inverse transformations to the full multivariate von Mises model concentration, circular mean and interaction matrix parametrization are then given by

$$\mu_i = 2 \arctan\left(\frac{\eta_i^s}{\eta_i^c + \sqrt{\eta_i^{c2} + \eta_i^{s2}}} \right) \quad \mod 2\pi,$$
$$\kappa_i = \sqrt{\eta_i^{c2} + \eta_i^{s2}}$$

where η_i^c and η_i^s are the exponential canonical parameters composing $\boldsymbol{\eta}$ in (B.5a) for $i = 1, \ldots, n$ associated with the cosine and sine respectively, and

$$\mathbf{D}^{cc} = 2\mathbf{E}^{cc} \odot c(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathsf{T}} + 2\mathbf{E}^{cs} \odot c(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathsf{T}} + 2\,\mathbf{E}^{cs\,\mathsf{T}} \odot s(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathsf{T}} + 2\mathbf{E}^{ss} \odot s(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathsf{T}}$$

$$\mathbf{D}^{cs} = 2\mathbf{E}^{cs} \odot c(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathsf{T}} - 2\mathbf{E}^{cc} \odot c(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathsf{T}} + 2\mathbf{E}^{ss} \odot s(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathsf{T}} - 2\,\mathbf{E}^{cs\,\mathsf{T}} \odot s(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathsf{T}}$$

$$\mathbf{D}^{ss} = 2\mathbf{E}^{ss} \odot c(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathsf{T}} - 2\mathbf{E}^{cs} \odot s(\boldsymbol{\mu})c(\boldsymbol{\mu})^{\mathsf{T}} - 2\,\mathbf{E}^{cs\,\mathsf{T}} \odot c(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathsf{T}} + 2\mathbf{E}^{cc} \odot s(\boldsymbol{\mu})s(\boldsymbol{\mu})^{\mathsf{T}}.$$

The multivariate von Mises model (B.2) and the full multivariate von Mises model (B.3) coincide when the precision matrix $\mathbf{D}$ is of the form

$$\mathbf{D} = \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{\Lambda} \end{bmatrix}.$$

In terms of differential geometry, the parametrizations constitute coordinate systems on the manifolds of probability measures. In this regard, the relationship between the dimensionality of the full multivariate von Mises statistical model S_{fMvM} given by (B.9) and the multivariate von Mises distribution S_{MvM} given by (B.2) is

$$\dim\left(S_{\text{fMvM}}\right) = 2n(n+1) > \frac{n(n+3)}{2} = \dim\left(S_{\text{MvM}}\right),$$

and as a consequence, $S_{\text{MvM}} \subset S_{\text{fMvM}}$ for $\boldsymbol{\kappa} > 0$, which implies that the multivariate von Mises statistical model is a submanifold of the full multivariate von Mises model.

B.2. The full multivariate von Mises submodel. To reduce the computational burden of sampling from a $2n\,(n+1)$ dimensional statistical model, where n denotes the dimensionality of the supporting torus, and to improve the stability of the Fisher metric estimate, we further reduce the full multivariate von Mises statistical model (B.9) dimensionality by removing interactions between cosines and sines for the same angle.

Considering the exponential rewriting of the full multivariate von Mises model (B.4) as a graphical interaction model [21], this is equivalent to removing direct feedback loops from the system. Mathematically, this means setting the elements e_{ij}^{cc}, e_{ij}^{ss}, e_{ij}^{cs} of $\mathbf{E}^{cc}$, $\mathbf{E}^{ss}$, $\mathbf{E}^{cs}$ submatrices of $\mathbf{E}$ in (B.5b) for $i = j$ equal to zero. As a consequence $\mathbf{D}^{cc}$, $\mathbf{D}^{ss}$, $\mathbf{D}^{cs}$ submatrices of the matrix $\mathbf{D}$ representing interactions in (B.9) have zero diagonals. The dimension of this specific submodel is $2n^2$.

B.3. Extended multivariate von Mises Gibbs sampler. To generate realizations from the (B.4) distributed random vectors for the computation of the natural gradient estimates (A.3), we use the multi-stage Gibbs sampler [61]. The univariate conditionals of the exponential rewriting of the extended multivariate von Mises distribution (B.4) are given by
(B.11)
$$f(\theta_k | \theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_n) \propto \exp\left\{\gamma_k^1 \cos(\theta_k - \nu_k^1) + \gamma_k^2 \cos(2(\theta_k - \nu_k^2))\right\},$$

where

$$\gamma_k^1 = \left\{ \left(\eta_k^c + \sum_{\substack{i=1 \\ i \neq k}}^{n}(e_{ki}^{cc} + e_{ik}^{cc})\cos(\theta_i) + \sum_{\substack{i=1 \\ i \neq k}}^{n}(e_{ki}^{cs} + e_{ik}^{sc})\sin(\theta_i) \right)^2 + \right.$$

$$\left. + \left(\eta_k^s + \sum_{\substack{i=1 \\ i \neq k}}^{n}(e_{ki}^{ss} + e_{ik}^{ss})\sin(\theta_i) + \sum_{\substack{i=1 \\ i \neq k}}^{n}(e_{ki}^{sc} + e_{ik}^{cs})\cos(\theta_i) \right)^2 \right\}^{\frac{1}{2}},$$

$$\nu_k^1 = \operatorname{atan2}\left(\frac{\eta_k^s + \sum_{\substack{i=1 \\ i \neq k}}^{n}(e_{ki}^{ss} + e_{ik}^{ss})\sin(\theta_i) + \sum_{\substack{i=1 \\ i \neq k}}^{n}(e_{ki}^{sc} + e_{ik}^{cs})\cos(\theta_i)}{\eta_k^c + \sum_{\substack{i=1 \\ i \neq k}}^{n}(e_{ki}^{cc} + e_{ik}^{cc})\cos(\theta_i) + \sum_{\substack{i=1 \\ i \neq k}}^{n}(e_{ki}^{cs} + e_{ik}^{sc})\sin(\theta_i)} \right),$$

$$\gamma_k^2 = \left\{ (e_{kk}^{cc} - e_{kk}^{ss})^2 + (e_{kk}^{sc} + e_{kk}^{cs})^2 \right\}^{\frac{1}{2}},$$

$$\nu_k^2 = \frac{1}{2}\operatorname{atan2}\left(\frac{e_{kk}^{sc} + e_{kk}^{cs}}{e_{kk}^{cc} - e_{kk}^{ss}} \right),$$

where $\operatorname{atan2}(\frac{y}{x})$ is the argument of the complex number $x + iy$, η_i^c and η_i^s are the canonical parameters in $\boldsymbol{\eta}$ associated with $\kappa_i \cos(\mu_i)$ and $\kappa_i \sin(\mu_i)$ in (B.5a) respectively, and e_{ij}^{cs}, e_{ij}^{sc}, e_{ij}^{cc}, e_{ij}^{ss} are the elements of the $\mathbf{E}^{cs}$, $\mathbf{E}^{sc}$, $\mathbf{E}^{cc}$, $\mathbf{E}^{cc}$ submatrices of $\mathbf{E}$ in (B.5b).

The probability density function (B.11) is of the generalized von Mises distribution introduced in [32]. We implement von Neumann's rejection sampling algorithm of [31] to generate samples from the generalized multivariate von Mises distribution.

Implementation of the extended multivariate von Mises Gibbs sampler is presented in Algorithm (B.1) in pseudocode form. The Gibbs sampler's number of iterations m was determined experimentally and is set to 100.

Algorithm B.1 Extended multivariate von Mises Gibbs sampler

Require: $\boldsymbol{\eta}$, n, m
Ensure: $\mathbf{U}_{d \times n} \sim \operatorname{EMvM}_d(\boldsymbol{\eta})$
  for $k = 1$ to $d - 1$ **do**
    $\mathbf{U}_{k,1:} \leftarrow$ Generate n i.i.d. $GvM_2\left(\sqrt{(\eta_k^c)^2 + (\eta_k^s)^2},\ \operatorname{atan2}\left(\frac{\eta_k^s}{\eta_k^c}\right),\ \gamma_k^2,\ \nu_k^2 \right)$
  end for
  for $j = 1$ to n **do**
    $\mathbf{U}_{d,j} \leftarrow GvM_2\left(\gamma_d^1,\ \nu_d^1,\ \gamma_d^2,\ \nu_d^2 | \mathbf{U}_{1,j}, \ldots, \mathbf{U}_{d-1,j} \right)$
  end for
  for $i = 1$ to m **do**
    for $k = 1$ to d **do**
      for $j = 1$ to n **do**
        $\mathbf{U}_{k,j} \leftarrow GvM_2\left(\gamma_k^1,\ \nu_k^1,\ \gamma_k^2,\ \nu_k^2 \big| \mathbf{U}_{1,j}, \ldots, \mathbf{U}_{k-1,j}, \mathbf{U}_{k+1,j}, \ldots, \mathbf{U}_{d,j} \right)$
      end for
    end for
  end for

Appendix C. Adaptive quantile hill climbing.

For fixed q_t-quantile, iteratively solving (3.1) with the expected fitness of the form (3.15) is equivalent to solving

$$\text{(C.1)} \qquad \boldsymbol{\theta}^{t+1} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \, p(\mathbf{x}|\boldsymbol{\theta}),$$

$$\text{(C.2)} \qquad \text{s.t. } \mathbf{x} \sim P(\boldsymbol{\theta}^t) \mid \mathbf{F}(\mathbf{x}) \geq \widehat{\mathbf{F}^{\boldsymbol{\theta}^t}_{1-\frac{1}{q_t}}}.$$

The update step $\boldsymbol{\theta}^{t+1}$ is given by maximizing the likelihood $p(\mathbf{x}|\boldsymbol{\theta})$ over (C.2). In this context, the adaptive selection quantile can be formulated as a more general search method, summarized in the pseudocode in Algorithm C.1.

Algorithm C.1 Adaptive quantile hill climbing

Require: parametric family $P(\boldsymbol{\theta})$, $\mathbf{F}(\mathbf{x})$, N, $q_{\max}$, β
Ensure: $\mathbf{x}_{opt}$
  repeat

$\qquad \cos(\alpha^{t-1}) \leftarrow \frac{<\Delta\boldsymbol{\theta}^{t-1}, \Delta\boldsymbol{\theta}^{t-2}>_{F(\boldsymbol{\theta}^{t-2})}}{||\Delta\boldsymbol{\theta}^{t-1}||_{F(\boldsymbol{\theta}^{t-2})}||\Delta\boldsymbol{\theta}^{t-2}||_{F(\boldsymbol{\theta}^{t-2})}}$

$\qquad q_t \leftarrow \min\left[q_{t-1} \exp\left\{ \beta \cos(\alpha^{t-1}) \right\}, q_{\max} \right]$
$\qquad$ Generate N i.i.d. samples $\mathbf{x}_i \sim P(\boldsymbol{\theta}^{t-1})$
$\qquad$ Assign ranks r_j to $\mathbf{F}(\mathbf{x}_i)$
$\qquad$ Update $\boldsymbol{\theta}^t$ by performing a maximum likelihood estimate of $P(\boldsymbol{\theta})$ from the ranked
$\qquad$ samples $\mathbf{x}_{r_N}, \mathbf{x}_{r_{N-1}}, \ldots, \mathbf{x}_{r_{N-\frac{N}{q_t}}}$
  until stopping criterium is met

Appendix D. Implementation details.

We present a few valid technical details and algorithmic settings here. These include the map between the unit flat n-torus and the optimization configuration space Appendix D.1, the penalty function used to integrate nonlinear constraints into the optimization schedule Appendix D.2, the non–overlapping constraint violation formulation Appendix D.3, the adaptive learning rate Appendix D.4, the Fisher metric tensor scaling used to stabilize the unit natural gradients Appendix D.5, the hyperparameter tuning method we developed Appendix D.6 and finally, the acceleration of computations through parallelization Appendix D.7.

D.1. Boundary mapping. The statistical model (B.3) we are working with consists of a probability distribution with its support on the n-torus T^n whose product space components are unit circles and is topologically equivalent to the identification space [86]

$$\prod_{i=1}^{n} [0, 2\pi] \Big/ {\sim}$$

where the equivalence relation $\sim$ is defined by identifying all points such that

$$(0, x_2, \ldots, x_n) \sim (2\pi, x_2, \ldots, x_n),$$
$$(x_1, 0, \ldots, x_n) \sim (x_1, 2\pi, \ldots, x_n),$$
$$\vdots$$
$$(x_1, x_2, \ldots, 0) \sim (x_1, x_2, \ldots, 2\pi).$$

We need to map the n-cube $[0, 2\pi]^n$ to an n-orthotope defined by the boundary constraints of the optimization problem in order to evaluate the fitness of each realization of the extended multivariate von Mises distributed random vector. Specifically, we define a map $y_i : [0, 2\pi) \to [l_i, u_i)$ via

$$y_i = \frac{x_i}{2\pi}(u_i - l_i) + l_i \ , \ x_i \in [0, 2\pi)$$

for $i = 1, \ldots, n$ where u_i and l_i are i-th upper and lower bound, respectively.

This kind of mapping is natural for variables with inherent periodicity but somewhat problematic for nonperiodic ones due to the discontinuity in the configuration space introduced by identifying lower and upper bounds, and more importantly, $\mathbf{F}(x_1, \ldots, l_a, \ldots, x_n) \neq \mathbf{F}(x_1, \ldots, u_a, \ldots, x_n)$ in the case when there is no period p such that $l_a = u_a \mod p$.

To address this inconvenience, we define a different boundary map for aperiodic variable x_a, given by

$$y_a = \begin{cases} l_a + \frac{x_a}{\pi}(u_a - l_a) & \text{if} \quad x_a \in [0, \pi), \\ 2u_a - l_a - \frac{x_i}{\pi}(u_a - l_a) & \text{if} \quad x_a \in [\pi, 2\pi), \end{cases}$$

at the expense of loosing injectivity of the boundary map $y_a : [0, 2\pi) \to [l_a; u_a)$ and introducing additional extrema in the optimization landscape. In practice, this is not a problem since the algorithm is built to be robust in complex optimization landscapes.

After combining the above, we have the following boundary constraint mapping

$$y_i = \begin{cases} l_i + \frac{x_i}{2\pi}(u_i - l_i) & \text{if} \quad x_i \text{ is periodic} \quad \wedge \quad x_i \in [0, 2\pi), \\ l_i + \frac{x_i}{\pi}(u_i - l_i) & \text{if} \quad x_i \text{ is aperiodic} \quad \wedge \quad x_i \in [0, \pi), \\ 2u_i - l_i - \frac{x_i}{\pi}(u_i - l_i) & \text{if} \quad x_i \text{ is aperiodic} \quad \wedge \quad x_i \in [\pi, 2\pi), \end{cases}$$

for $i = 1, \ldots, n$.

D.2. Constraint handling. To address linear and nonlinear constraints, we implement a penalty function based on feasibility considerations [25]. The basic premise is to create an ordering on the set of candidate solutions, such that: i) any feasible solution is better than any infeasible one, ii) between two feasible solutions, the one with better objective function is preferred, and iii) between two infeasible solutions, the one with lower constraint violation is preferred. Then for the following minimization problem

(D.1) $$\min_{\mathbf{x} \in \mathbf{x}} f(\mathbf{x}),$$

$$\text{s.t. } g_j(\mathbf{x}) \leq 0, \ j = 1, \ldots, J,$$

where $g_j(\mathbf{x})$ are inequality constraints, given a set of solutions $\{\mathbf{x}_i \mid i = 1, \ldots, N\}$, the penalty function is expressed as

(D.2) $$\mathbf{F}(\mathbf{x}_i) = \begin{cases} f(\mathbf{x}_i) & \text{if} \ g_j(\mathbf{x}_i) \leq 0, \forall \ j = 1, \ldots, J \\ f^{\max} + \sum_{k \in K} \frac{g_k(\mathbf{x}_i)}{g_k^{\max}} & \text{otherwise} \end{cases}$$

where $K = \{j : g_j(\mathbf{x}_i) > 0, \ j = 1, \ldots, J\}$ and

(D.3) $$f^{\max} = \max\{f(\mathbf{x}_i) \mid \mathbf{x}_i : \ g_j(\mathbf{x}_i) \leq 0, \ i = 1, \ldots, N \ , \ j = 1, \ldots, J\},$$

(D.4) $$g_j^{\max} = \max\{g_j(\mathbf{x}_i) \mid \mathbf{x}_i : \ g_j(\mathbf{x}_i) > 0, \ i = 1, \ldots, N\}.$$

Note that this is particularly well suited for the adaptive selection quantile introduced in subsection 3.2 since only the ordering is considered for the trust region updates. Additionally, the penalty function (D.2) can be easily augmented for multiple objectives by using some suitable aggregation function $G(f_1(\mathbf{x}), \ldots, f_R(\mathbf{x}))$ where $f_r(\mathbf{x})$ for $r = 1, \ldots, R$ are the multiple objective functions [20].

D.3. G-packing overlap constraint evaluation of convex polytopes. To evaluate the intersection and the degree of constraint violation between convex polytopes in candidate solutions, we use the method based on Phi-functions [19]. Given the convex polytope

$$P_0 = \mathrm{conv}\{V_0 | V_0 = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m), \mathbf{p}_i \in \mathbb{R}^n\},$$

centred at the origin and defined by the convex hull of the vertices $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_m$, and given rotated and translated copies of P_0

$$P_1 = \mathrm{conv}\{V_1 | V_1 = R_1 V_0^T + c_1\},$$
$$P_2 = \mathrm{conv}\{V_2 | V_2 = R_2 V_0^T + c_2\},$$

for some rotation matrices R_1, R_2 and translation vectors c_1, c_2, Minkowsky's separating hyperplane theorem [39], states that P_1 and P_2 do not intersect if and only if there exists a separating hyperplane parallel to at least one of the faces of P_1 or P_2. It means that the hyperplanes defined as extensions of faces of P_1 and P_2 are all candidate separating hyperplanes, and it is adequate to check vertices of P_1 against P_2 hyperplanes and vice versa.

To implement this, we set hyperplanes defining polygon P_0 to

(D.5)
$$
\begin{array}{ccccccccc}
a_{11}x_1 & + & \ldots & + & a_{1n}x_n & + & b_1 & = & 0 \\
a_{21}x_1 & + & \ldots & + & a_{2n}x_n & + & b_2 & = & 0 \\
\vdots & & & & \vdots & & \vdots & & \vdots \\
a_{m1}x_1 & + & \ldots & + & a_{mn}x_n & + & b_m & = & 0
\end{array}
$$

where $\mathbf{a}_i$ are the unit normal vectors of the hyperplanes defined by the faces of the polytope pointing outward from the polygon and express vertices of P_1 in the coordinate system of P_2 denoted by V_{12} and vertices of P_2 in the coordinate system of P_1 denoted by V_{21} as

$$V_{12} = R_2^{-1}[V_1^T - c_2],$$
$$V_{21} = R_1^{-1}[V_2^T - c_1].$$

Since $\|\mathbf{a}_i\| = 1$ by inserting V_{12} and V_{21} into the hyperplane equations (D.5) we not only check for the existence of a separating hyperplane but in practice compute the euclidean distance between P_1 and P_2 which has the closed form expression

(D.6) $$\mathrm{dist}(P_1, P_2) = \max\{\max_i \min_{\mathbf{x} \in V_{12}} \mathbf{a}_i^T \mathbf{x} + b_i, \max_i \min_{\mathbf{x} \in V_{21}} \mathbf{a}_i^T \mathbf{x} + b_i\},$$

for $i = 1, \ldots, m$. The distance function (D.6) is a continuous and piecewise differentiable function, and P_1 and P_2 do not intersect if and only if $\mathrm{dist}(P_1, P_2) \geq 0$.

To evaluate whether a collection of convex polytopes $\mathcal{K}$ created by a CSG G acting on a convex polytope P_0 is a G-packing, for $\mathcal{K}_G$ in (2.2) we define

(D.7) $$\mathrm{dist}(\mathcal{K}_G) = \min_{i,j,\{\alpha\}} \mathrm{dist}\left(g_i P_0, g_{j\{\alpha\}} P_0\right),$$

where $g_i,\ g_{j\{\alpha\}} \in G$ such that

(D.8)
$$g_i P_0 = R_i P_0 + \mathbf{a}_i,$$

(D.9)
$$g_{j\{\alpha\}} P_0 = R_j P_0 + \mathbf{a}_j + \mathbf{l}_{\{\alpha\}},$$

for $i, j = 1, \ldots, N$, $\{\alpha\} = \{(u_1, \ldots, u_n) | u_i \in \{-2, -1, 0, 1, 2\}\}$ and $i \neq j$ if $\{\alpha\} = (0, \ldots, 0)$ where $\mathbf{l}_{\{\alpha\}} \in L$ (2.1) and $\mathbf{a}_i$ is of the form (2.4). In other words, we compute minimal euclidean distances between orbits of P_0 (D.8) whose centroids lie inside the primitive cell and orbits of P_0 (D.9) whose centroids lie inside neighbouring primitive cells to up to twice the lattice basis vectors $\mathbf{b}_1, \ldots, \mathbf{b}_n$.

The requirement that the set of possible coefficients of lattice element $\mathbf{l}_{\{\alpha\}}$ is restricted to $\{-2, -1, 0, 1, 2\}$ stems from the observation that if the upper bound on the size of the lattice vector generators is set to $\|\mathbf{b}_i\| = Nd$, where d is the diameter of circumscribed $n - 1$-sphere of P_0 and N is the number of symmetry operations in G, then to evaluate whether $\mathcal{K}_G$ is a G-packing it is adequate to assess the intersection between $g_i P_0$ and $g_{j\{\alpha\}} P_0$ for $\mathbf{l}_{\{\alpha\}} \in L$ up to twice the lattice basis vectors $\mathbf{b}_1, \ldots, \mathbf{b}_n$ in every coordinate direction.

Based on (D.7) $\mathcal{K}_G$ is a plane group packing if and only if dist $(\mathcal{K}_G) \geq 0$. the In case of dist $(\mathcal{K}_G) < 0$, we get a measure of constraint violation of a candidate solution defined by $|\text{dist}\,(\mathcal{K}_G)|$ used in penalty function (D.2) by setting $g\,(\mathbf{x}) = -\text{dist}\,(\mathcal{C}_G(\mathbf{x}))$ where $\mathcal{C}_G(\mathbf{x})$ is a $G \in \mathcal{G}$ configuration of the form (2.2a) defined by the candidate solution $\mathbf{x}$ that is not necessarily a packing.

D.4. Learning rates. Experiments show that having a single trust region radius for the exponential multivariate von Mises statistical manifold is insufficient and results in poor performance. Instead, we decide to transfer the unit gradient (3.8) back to the original circular mean, concentration and angle interaction parametrizations and perform the gradient ascent updates in those coordinates, allowing us to use an additional separate learning rate for each parameter group. In Appendix D.4.1, we introduce the aforementioned change of coordinates of the natural gradients. Additionally, we modify the adaptive learning rate scheme proposed in [69] described in Appendix D.4.2.

D.4.1. Circular mean, concentration and precision update equations. By differentiating (B.5) with respect to time, we get a system of following linear equations

$$\frac{d\eta_i^c}{dt} = -\kappa_i \sin(\mu_i) \frac{d\mu_i}{dt} + \cos(\mu_i) \frac{d\kappa_i}{dt},$$

$$\frac{d\eta_i^s}{dt} = \kappa_i \cos(\mu_i) \frac{d\mu_i}{dt} + \sin(\mu_i) \frac{d\kappa_i}{dt},$$

$$\begin{aligned}
\frac{de_{ij}^{cc}}{dt} &= \cos(\mu_i)\cos(\mu_j)\frac{dd_{ij}^{cc}}{dt} - \cos(\mu_i)\sin(\mu_j)\frac{dd_{ij}^{cs}}{dt} - \sin(\mu_i)\cos(\mu_j)\frac{dd_{ij}^{sc}}{dt} \\
&+ \sin(\mu_i)\sin(\mu_j)\frac{dd_{ij}^{ss}}{dt} - d_{ij}^{cs}\cos(\mu_i)\cos(\mu_j)\frac{d\mu_j}{dt} - d_{ij}^{sc}\cos(\mu_i)\cos(\mu_i)\frac{d\mu_i}{dt} \\
&- d_{ij}^{cc}\cos(\mu_j)\sin(\mu_i)\frac{d\mu_i}{dt} - d_{ij}^{cc}\cos(\mu_i)\sin(\mu_j)\frac{d\mu_j}{dt} + d_{ij}^{ss}\cos(\mu_i)\sin(\mu_j)\frac{d\mu_i}{dt} \\
&+ d_{ij}^{ss}\cos(\mu_j)\sin(\mu_i)\frac{d\mu_j}{dt} + d_{ij}^{cs}\sin(\mu_i)\sin(\mu_j)\frac{d\mu_i}{dt} + d_{ij}^{sc}\sin(\mu_i)\sin(\mu_j)\frac{d\mu_j}{dt},
\end{aligned}$$

$$\frac{de_{ij}^{cs}}{dt} = \cos(\mu_i)\cos(\mu_j)\frac{dd_{ij}^{cs}}{dt} + \cos(\mu_i)\sin(\mu_j)\frac{dd_{ij}^{cc}}{dt} - \sin(\mu_i)\cos(\mu_j)\frac{dd_{ij}^{ss}}{dt}$$
$$- \sin(\mu_i)\sin(\mu_j)\frac{dd_{ij}^{sc}}{dt} + d_{ij}^{cc}\cos(\mu_i)\cos(\mu_j)\frac{d\mu_j}{dt} - d_{ij}^{ss}\cos(\mu_i)\cos(\mu_j)\frac{d\mu_i}{dt}$$
$$- d_{ij}^{cs}\cos(\mu_j)\sin(\mu_i)\frac{d\mu_i}{dt} - d_{ij}^{cs}\cos(\mu_i)\sin(\mu_j)\frac{d\mu_j}{dt} - d_{ij}^{sc}\cos(\mu_i)\sin(\mu_j)\frac{d\mu_i}{dt}$$
$$- d_{ij}^{sc}\cos(\mu_j)\sin(\mu_i)\frac{d\mu_j}{dt} - d_{ij}^{cc}\sin(\mu_i)\sin(\mu_j)\frac{d\mu_i}{dt} + d_{ij}^{ss}\sin(\mu_i)\sin(\mu_j)\frac{d\mu_j}{dt},$$

$$\frac{de_{ij}^{sc}}{dt} = \cos(\mu_i)\cos(\mu_j)\frac{dd_{ij}^{sc}}{dt} + \cos(\mu_j)\sin(\mu_i)\frac{dd_{ij}^{cc}}{dt} - \cos(\mu_i)\sin(\mu_j)\frac{dd_{ij}^{ss}}{dt}$$
$$- \sin(\mu_i)\sin(\mu_j)\frac{dd_{ij}^{cs}}{dt} + d_{ij}^{cc}\cos(\mu_i)\cos(\mu_j)\frac{d\mu_i}{dt} - d_{ij}^{ss}\cos(\mu_i)\cos(\mu_j)\frac{d\mu_j}{dt}$$
$$- d_{ij}^{cs}\cos(\mu_i)\sin(\mu_j)\frac{d\mu_i}{dt} - d_{ij}^{cs}\cos(\mu_j)\sin(\mu_i)\frac{d\mu_j}{dt} - d_{ij}^{sc}\cos(\mu_j)\sin(\mu_i)\frac{d\mu_i}{dt}$$
$$- d_{ij}^{sc}\cos(\mu_i)\sin(\mu_j)\frac{d\mu_j}{dt} - d_{ij}^{cc}\sin(\mu_i)\sin(\mu_j)\frac{d\mu_j}{dt} + d_{ij}^{ss}\sin(\mu_i)\sin(\mu_j)\frac{d\mu_i}{dt},$$

$$\frac{de_{ij}^{ss}}{dt} = \cos(\mu_i)\cos(\mu_j)\frac{dd_{ij}^{ss}}{dt} + \cos(\mu_j)\sin(\mu_i)\frac{dd_{ij}^{cs}}{dt} + \cos(\mu_i)\sin(\mu_j)\frac{dd_{ij}^{sc}}{dt}$$
$$+ \sin(\mu_i)\sin(\mu_j)\frac{dd_{ij}^{cc}}{dt} + d_{ij}^{cs}\cos(\mu_i)\cos(\mu_j)\frac{d\mu_i}{dt} + d_{ij}^{sc}\cos(\mu_i)\cos(\mu_j)\frac{d\mu_j}{dt}$$
$$+ d_{ij}^{cc}\cos(\mu_i)\sin(\mu_j)\frac{d\mu_i}{dt} + d_{ij}^{cc}\cos(\mu_j)\sin(\mu_i)\frac{d\mu_j}{dt} - d_{ij}^{ss}\cos(\mu_j)\sin(\mu_i)\frac{d\mu_i}{dt}$$
$$- d_{ij}^{ss}\cos(\mu_i)\sin(\mu_j)\frac{d\mu_j}{dt} - d_{ij}^{cs}\sin(\mu_i)\sin(\mu_j)\frac{d\mu_j}{dt} - d_{ij}^{sc}\sin(\mu_i)\sin(\mu_j)\frac{d\mu_i}{dt}.$$

Consequently, by solving this system, we express the circular mean, concentration and precision parameter time derivatives in terms of canonical exponential parameters time derivatives. Then, the coordinate change between these two tangent spaces is

$$(\text{D.10a}) \qquad \frac{d\mu_i}{dt} = -\frac{\sin(\mu_i)}{\kappa_i}\frac{d\eta_i^c}{dt} + \frac{\cos(\mu_i)}{\kappa_i}\frac{d\eta_i^s}{dt},$$

$$(\text{D.10b}) \qquad \frac{d\kappa_i}{dt} = \cos(\mu_i)\frac{d\eta_i^c}{dt} + \sin(\mu_i)\frac{d\eta_i^s}{dt},$$

$$(\text{D.10c}) \quad \frac{dd_{ij}^{cc}}{dt} = d_{ij}^{cs}\frac{d\mu_j}{dt} + d_{ij}^{sc}\frac{d\mu_i}{dt} + \cos(\mu_i)\cos(\mu_j)\frac{de_{ij}^{cc}}{dt} + \cos(\mu_i)\sin(\mu_j)\frac{de_{ij}^{cs}}{dt}$$
$$+ \cos(\mu_j)\sin(\mu_i)\frac{de_{ij}^{sc}}{dt} + \sin(\mu_i)\sin(\mu_j)\frac{de_{ij}^{ss}}{dt},$$

$$(\text{D.10d}) \quad \frac{dd_{ij}^{cs}}{dt} = d_{ij}^{ss}\frac{d\mu_i}{dt} - d_{ij}^{cc}\frac{d\mu_j}{dt} + \cos(\mu_i)\cos(\mu_j)\frac{de_{ij}^{cs}}{dt} - \cos(\mu_i)\sin(\mu_j)\frac{de_{ij}^{cc}}{dt}$$
$$+ \cos(\mu_j)\sin(\mu_i)\frac{de_{ij}^{ss}}{dt} - \sin(\mu_i)\sin(\mu_j)\frac{de_{ij}^{sc}}{dt},$$

(D.10e) $\quad \dfrac{dd_{ij}^{sc}}{dt} = d_{ij}^{ss}\dfrac{d\mu_j}{dt} - d_{ij}^{cc}\dfrac{d\mu_i}{dt} + \cos(\mu_i)\cos(\mu_j)\dfrac{de_{ij}^{sc}}{dt} - \cos(\mu_j)\sin(\mu_i)\dfrac{de_{ij}^{cc}}{dt}$

$$+ \cos(\mu_i)\sin(\mu_j)\frac{de_{ij}^{ss}}{dt} - \sin(\mu_i)\sin(\mu_j)\frac{de_{ij}^{cs}}{dt},$$

(D.10f) $\quad \dfrac{dd_{ij}^{ss}}{dt} = -d_{ij}^{sc}\dfrac{d\mu_j}{dt} - d_{ij}^{cs}\dfrac{d\mu_i}{dt} + \cos(\mu_i)\cos(\mu_j)\dfrac{de_{ij}^{ss}}{dt} - \cos(\mu_j)\sin(\mu_i)\dfrac{de_{ij}^{cs}}{dt}$

$$- \cos(\mu_i)\sin(\mu_j)\frac{de_{ij}^{sc}}{dt} + \sin(\mu_i)\sin(\mu_j)\frac{de_{ij}^{cc}}{dt},$$

for $i, j = 1, \ldots, n$ and $\boldsymbol{\kappa} > 0$ where η_i^c and η_i^s are canonical parameters associated with $\kappa_i\cos(\mu_i)$ in (B.5a) and $\kappa_i\sin(\mu_i)$ respectively, e_{ij}^{cs}, e_{ij}^{sc}, e_{ij}^{cc}, e_{ij}^{ss} are elements of the $\mathbf{E}^{cs}$, $\mathbf{E}^{sc}$, $\mathbf{E}^{cc}$, $\mathbf{E}^{ss}$ submatrices of $\mathbf{E}$ in (B.5b) and d_{ij}^{cs}, d_{ij}^{sc}, d_{ij}^{cc}, d_{ij}^{ss} are elements of the $\mathbf{D}^{cs}$, $\mathbf{D}^{sc}$, $\mathbf{D}^{cc}$, $\mathbf{D}^{cc}$ submatrices of of the precision matrix $\mathbf{D}$ in (B.6).

Using the canonical parametrization time derivatives given by the flow associated with the natural gradient Monte-Carlo estimates (A.3)

$$\frac{d\boldsymbol{\eta}}{dt} = \widetilde{\nabla}_{\boldsymbol{\eta}}\widehat{J(\boldsymbol{\eta}, \mathbf{E})},$$
$$\frac{d\mathbf{E}}{dt} = \widetilde{\nabla}_{\mathbf{E}}\widehat{J(\boldsymbol{\eta}, \mathbf{E})},$$

and using the coordinate transformations (D.10), provide a flow associated with the circular mean, concentration and interaction parameters

$$\frac{d\boldsymbol{\mu}}{dt} = \widetilde{\nabla}_{\boldsymbol{\mu}}\widehat{J(\boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{D})},$$
$$\frac{d\boldsymbol{\kappa}}{dt} = \widetilde{\nabla}_{\boldsymbol{\kappa}}\widehat{J(\boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{D})},$$
$$\frac{d\mathbf{D}}{dt} = \widetilde{\nabla}_{\mathbf{D}}\widehat{J(\boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{D})}.$$

Thus the updated equations are then given by

(D.11a) $$\boldsymbol{\mu}^{t+1} = \boldsymbol{\mu}^t + \gamma_{\boldsymbol{\mu}}\widetilde{\nabla}_{\boldsymbol{\mu}}\widehat{J(\boldsymbol{\mu}^t, \boldsymbol{\kappa}^t, \mathbf{D}^t)},$$

(D.11b) $$\boldsymbol{\kappa}^{t+1} = \boldsymbol{\kappa}^t + \gamma_{\boldsymbol{\kappa}}\widetilde{\nabla}_{\boldsymbol{\kappa}}\widehat{J(\boldsymbol{\mu}^t, \boldsymbol{\kappa}^t, \mathbf{D}^t)},$$

(D.11c) $$\mathbf{D}^{t+1} = \mathbf{D}^t + \gamma_{\mathbf{D}}\widetilde{\nabla}_{\mathbf{D}}\widehat{J(\boldsymbol{\mu}^t, \boldsymbol{\kappa}^t, \mathbf{D}^t)},$$

where $\gamma_{\boldsymbol{\mu}}$, $\gamma_{\boldsymbol{\kappa}}, \gamma_{\mathbf{D}}$ are the respective learning rates for each parameter group.

D.4.2. Adaptive learning rates. To further stabilize the dynamical system given (D.11), we implement a method to adaptively adjust learning rates proposed in [69] by changing each variable's learning rate individually by comparing gradient directions at two consecutive steps. If the general trend is the same, the learning rate is increased to accelerate convergence. If the change in the path is significant, the learning rate is decreased to allow steps with smaller granularity.

In our setting, instead of comparing gradients directly, we compare parameter update differences given by

$$\Delta\mu_i^t = \mu_i^t - \mu_i^{t-1},$$
$$\Delta\kappa_i^t = \kappa_i^t - \kappa_i^{t-1},$$
$$\Delta d_{ij}^t = d_{ij}^t - d_{ij}^{t-1},$$

at times t and $t-1$ for $i,j = 1,\ldots,n$, and we set the upper bound for the adaptive learning rates to the initial learning rate γ^0 since we want only more fine-tuned learning rates when the algorithm has already located an optimum basin. The adaptive learning rate update equations then take the following form

(D.12a)
$$\gamma_{\mu_i}^t = \begin{cases} \min\left\{c_{up}\gamma_{\mu_i}^{t-1},\ \gamma_{\mu_i}^0\right\} & \text{if} \quad \text{sgn}\left(\Delta\mu_i^t\right) = \text{sgn}\left(\Delta\mu_i^{t-1}\right), \\ c_{down}\gamma_{\mu_i}^{t-1} & \text{if} \quad \text{sgn}\left(\Delta\mu_i^t\right) \neq \text{sgn}\left(\Delta\mu_i^{t-1}\right), \end{cases}$$

(D.12b)
$$\gamma_{\kappa_i}^t = \begin{cases} \min\left\{c_{up}\gamma_{\kappa_i}^{t-1},\ \gamma_{\kappa_i}^0\right\} & \text{if} \quad \text{sgn}\left(\Delta\kappa_i^t\right) = \text{sgn}\left(\Delta\kappa_i^{t-1}\right), \\ c_{down}\gamma_{\kappa_i}^{t-1} & \text{if} \quad \text{sgn}\left(\Delta\kappa_i^t\right) \neq \text{sgn}\left(\Delta\kappa_i^{t-1}\right), \end{cases}$$

(D.12c)
$$\gamma_{d_{ij}}^t = \begin{cases} \min\left\{c_{up}\gamma_{d_{ij}}^{t-1},\ \gamma_{d_{ij}}^0\right\} & \text{if} \quad \text{sgn}\left(\Delta d_{ij}^t\right) = \text{sgn}\left(\Delta d_{ij}^{t-1}\right), \\ c_{down}\gamma_{d_{ij}}^{t-1} & \text{if} \quad \text{sgn}\left(\Delta d_{ij}^t\right) \neq \text{sgn}\left(\Delta d_{ij}^{t-1}\right), \end{cases}$$

for $i,j = 1,\ldots,n$ where $\text{sign}(\cdot)$ is the signum function and $c_{up} > 1$, $c_{down} < 1$ are real positive constants.

Additionally, we use momentum constants $\alpha_{\boldsymbol{\mu}} > 0$, $\alpha_{\boldsymbol{\kappa}} > 0$ and $\alpha_{\mathbf{D}} > 0$ in the update equations by setting

$$m_{\mu_i}^t = \widetilde{\nabla}_{\mu_i} J\widehat{(\boldsymbol{\mu}^t, \boldsymbol{\kappa}^t}, \mathbf{D}^t) + \alpha_{\boldsymbol{\mu}} m_{\mu_i}^{t-1},$$
$$m_{\kappa_i}^t = \widetilde{\nabla}_{\kappa_i} J\widehat{(\boldsymbol{\mu}^t, \boldsymbol{\kappa}^t}, \mathbf{D}^t) + \alpha_{\boldsymbol{\kappa}} m_{\kappa_i}^{t-1},$$
$$m_{d_{ij}}^t = \widetilde{\nabla}_{d_{ij}} J\widehat{(\boldsymbol{\mu}^t, \boldsymbol{\kappa}^t}, \mathbf{D}^t) + \alpha_{\mathbf{D}} m_{d_{ij}}^{t-1},$$

to further aid the trajectory stabilization with the final form of update equations being

$$\mu_i^{t+1} = \mu_i^t + \gamma_{\mu_i}^t m_{\mu_i}^t,$$
$$\kappa_i^{t+1} = \kappa_i^t + \gamma_{\kappa_i}^t m_{\kappa_i}^t,$$
$$d_{ij}^{t+1} = d_{ij}^t + \gamma_{d_{ij}}^t m_{d_{ij}}^t.$$

D.5. Fisher metric spectral radius scaling. Although the sample covariance matrix estimator used in the estimation of the Fisher metric expressed in the canonical exponential parameters in (A.3) is considered efficient and unbiased, this may not be the case when the intrinsic geometry of the space of positive definite symmetric matrices is taken in account [78], possibly resulting in instabilities of the entropic trust region updates with subsequent large fluctuations in the parameter trajectories when the number of samples is not adequately large.

To improve the stability of the gradients in (3.9), we set a new metric tensor by scaling the inverse of the Fisher matrix by its spectral radius according to

$$\mathcal{I}_{\boldsymbol{\theta}}^{\lambda_{\min}} = \frac{1}{\lambda_{\min}} \mathcal{I}_{\boldsymbol{\theta}},$$

where $\lambda_{\min}$ is the smallest eigenvalue of $\mathcal{I}_{\boldsymbol{\theta}}$ or the inverse of the spectral radius of $\mathcal{I}_{\boldsymbol{\theta}}^{-1}$. Then the entropic trust region step size (3.7) takes the form

(D.13)
$$\delta\boldsymbol{\theta}^t = \Delta^t \frac{\left(\mathcal{I}_{\boldsymbol{\theta}^t}^{\lambda_{\min}}\right)^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t)}{\sqrt{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t)^T \left(\mathcal{I}_{\boldsymbol{\theta}^t}^{\lambda_{\min}}\right)^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t)}}.$$

The spectral radius scaling can be regarded as an adaptive trust region radius Δ^t in (D.13) since

$$\frac{\left(\mathcal{I}_{\boldsymbol{\theta}^t}^{\lambda_{\min}}\right)^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t)}{\sqrt{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t)^T \left(\mathcal{I}_{\boldsymbol{\theta}^t}^{\lambda_{\min}}\right)^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t)}} = \sqrt{\lambda_{\min}} \frac{(\mathcal{I}_{\boldsymbol{\theta}^t})^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t)}{\sqrt{\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t)^T (\mathcal{I}_{\boldsymbol{\theta}^t})^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}^t)}},$$

Finally, the entropic trust region update equations (3.9) change to

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \sqrt{\lambda_{\min}} \frac{\widetilde{\nabla} J(\boldsymbol{\theta}^t)}{\| \widetilde{\nabla} J(\boldsymbol{\theta}^t) \|}.$$

D.6. Hyperparameters. Multiple hyperparameters need to be set for the optimization algorithm to work accurately. In the following few paragraphs, we address the specific values we use for the problem of dense plane group packings of convex polygons. All the values were determined experimentally on the problem of densest $p2$-packing of a regular octagon, described in more detail in section 4.

Initially, the number of samples used Monte–Carlo estimates of the Fisher metric tensor (A.1) and the expected fitness gradient (3.22) was set. As was already mentioned in section D.5, the sample covariance matrix estimator is known to be ill-conditioned when the ratio $c = \frac{p}{N}$ (where p is the rank of the covariance matrix and N the number of samples) is not negligible. Since the most computationally demanding part of the optimization schedule is the extended multivariate von Mises Gibbs sampler Appendix B.3, the trade-off between the accuracy and stability of the algorithm's trajectories and speed of computations has to be balanced. We experimentally determined $c = 0.12$ to be an acceptable balance between speed and stability.

Concerning the adaptive selection quantile parameters in (3.24), we set the starting q-quantile to $q_0 = \frac{N}{100}$ and the schedule parameter β in such a way that if after each iteration the cosine of the angle between consecutive steps is 0, corresponding to no change in the direction of the updates, the adaptive selection q-quantile would reach the Nth quantile after $2,000$ iterations; that is, $\beta = \ln\left(\frac{N}{q_0}\right) /2,000$.

Hyperparameter settings related to the adaptive learning rates described in section Appendix D.4.2 were obtained by a localized instance of grid search. First, the rate of change of the learning rates in (D.12) was set to $c_{up} = 1.1$ and $c_{down} = 0.9$. Subsequently, we constructed the 6D grid

$$\left\{\left(\gamma_{\boldsymbol{\mu}}^0, \gamma_{\boldsymbol{\kappa}}^0, \gamma_{\mathbf{D}}^0\right) | \gamma_*^0 \in \{0.25; 0.5; 0.75; 1\}\right\} \times \left\{\left(\alpha_{\boldsymbol{\mu}}^0, \alpha_{\boldsymbol{\kappa}}^0, \alpha_{\mathbf{D}}^0\right) | \alpha_*^0 \in \{0; 0.25; 0.5; 0.75\}\right\},$$

with $* \in \{\boldsymbol{\mu}, \boldsymbol{\kappa}, \mathbf{D}\}$, where γ_*^0 denotes the learning rate and α_*^0 momentum parameters associated with the extended multivariate von Mises distribution parameters in (B.3) and performed 20 optimization runs for each grid node. After the initial search, we increased the granularity of the grid by halving the size of the line segment between two neighbouring nodes and performed 20 optimization runs but only for the nearest neighbours of the node that attained the highest mean from the objective function of the best solutions found in each run in the initial grid search. We iterated this process of exploring only the nearest neighbour of the node that attained the best mean from all 20 runs in the previous iteration. When no better combination of parameters was found, we increased the granularity of the grid by halving the size of the grid line segments and repeated the exploration process on a more fine-grained grid, starting with the node with the highest mean of the objective.

$\gamma_{\boldsymbol{\mu}}^0$	0.140625
$\gamma_{\boldsymbol{\kappa}}^0$	0.171875
$\gamma_{\mathbf{D}}^0$	0.21875
c_{up}	1.1
c_{down}	0.9
$\alpha_{\boldsymbol{\mu}}$	0.7109375
$\alpha_{\boldsymbol{\kappa}}$	0.1953125
$\alpha_{\mathbf{D}}$	0.578125
q_0	$\frac{N}{100}$
β	$\ln\left(\frac{N}{q_0}\right)/2,000$
c	0.12
Number of iterations	8,000
c_ϵ	1.2

Table 2: Used hyperparameter settings.

The exponential decay c_ϵ of the $\boldsymbol{\epsilon}^r$ neighbourhood in (3.48) used in the refining part of the algorithm has to be balanced between accuracy and speed. If c_ϵ is set too high, the neighbourhoods converge too fast, and it is possible that some s the $\boldsymbol{\epsilon}^r$ neighbourhoods of the best solution found at rth run do not contain the optimal solution for all $r \geq s$. On the other hand, if c_ϵ is too low, it increases the number of refining runs and impacts the algorithm's efficiency. c_ϵ was set to 1.2 by evaluating the difference from the theoretical packing density (4.2) with the objective of accuracy being $\Delta_{\mathcal{K}_{p2}} < 10^{-7}$.

All the values of the hyperparameters used in computations presented in the experimental results section 4 and Appendix G are listed in Table 2.

D.7. Parallelizing computations. All the computations were performed in MATLAB R2021b. To accelerate computations, we utilize the architecture of modern microprocessors using the Parallel Computing toolbox. In the following paragraphs, we provide details on our parallel algorithm implementation.

Gibbs sampling for the exponential rewriting of the extended multivariate von Mises distribution described in Appendix B.3 is suitable for parallel computations on the CPU since the Gibbs sampling simulation involves evaluating scalar operations, non–elementary function calls and branching. We divide the number of samples λ needed for the trust region step size estimation (3.7) among W number of workers available. Then every worker generates $\frac{\lambda}{T}$ independent extended multivariate von Mises with parameter $(\boldsymbol{\eta}^t, \mathbf{E}^t)$ distributed samples.

After workers complete the Gibbs sampling simulations, the results are combined and transferred to the GPU memory, and all subsequent computations are done on the GPU. GPU computations are particularly efficient for the objective function (2.5), and constraint violation (D.7) evaluations since both involve just elementary function evaluations that are utilized using MATLAB's inherent vectorization design and are efficiently computed on the GPU for all λ candidate solutions in parallel.

Appendix E. Entropic trust region trajectories. A visualization of the evolution of the distribution parameters is presented in Figure 14 and Figure 15. As we already mention in Appendix B, it is generally difficult to precisely interpret the distribution parameters. Nevertheless, some intuition about the algorithmic behaviour can be extracted from the entropic trust region trajectories.
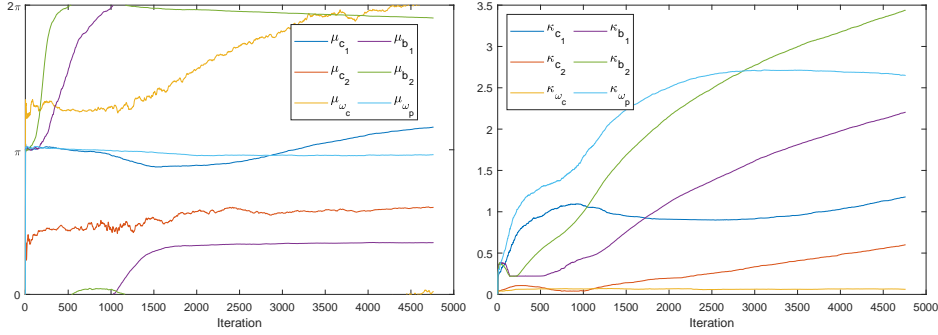
Figure 14: Trajectories of the extended multivariate von Mises distribution (left) circular mean $\boldsymbol{\mu}$ and (right) concentration $\boldsymbol{\kappa}$ parameters.
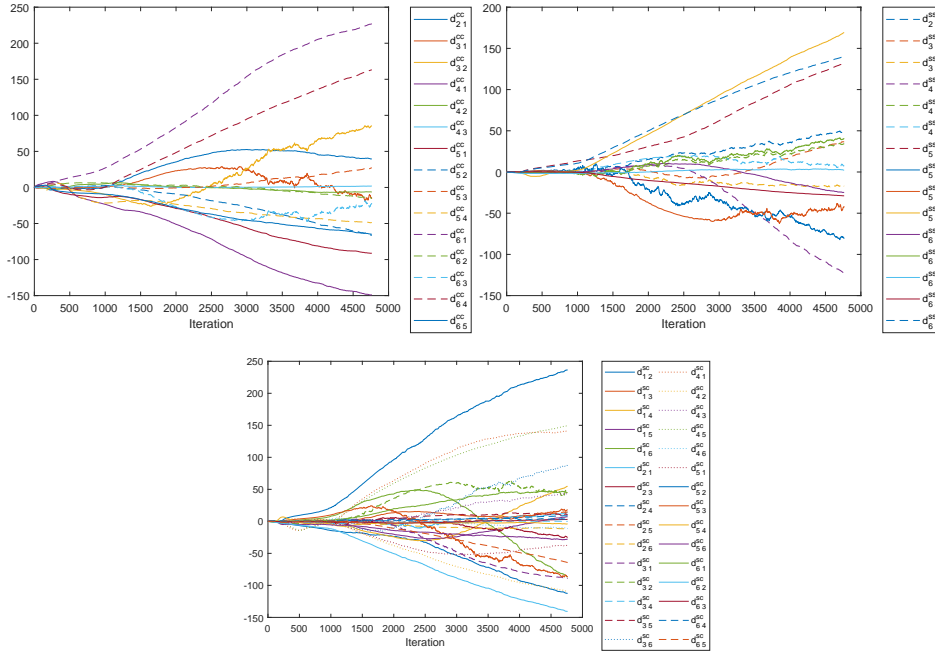

Figure 15: Trajectories of the extended multivariate von Mises distribution $\mathbf{D}$ matrix elements: (top left) $\mathbf{D}^{cc}$ representing $\cos(\theta_i) - \cos(\theta_j)$ interactions, (top right) elements of $\mathbf{D}^{ss}$ representing $\sin(\theta_i) - \sin(\theta_j)$ interactions and (bottom) elements of $\mathbf{D}^{sc}$ representing $\sin(\theta_i) - \cos(\theta_j)$ interactions.

Most of the circular mean $\boldsymbol{\mu}$ parameters stabilize after $3,000$ iterations of the algorithm with the exception of the parameter μ_{ω_c} which is related to the angle of rotation of the octagon in the asymmetric unit. This is due to the rotational symmetry of the octagon that induces multiple global optima in the ω_c subspace of the optimization landscape. To be precise, for $\omega_c = \frac{2\pi(k-1)}{8}$, $k = 1, \ldots, 7$ the packing density (2.5) is equal.

Similar behaviour can be observed in the trajectories of the concentration parameters $\boldsymbol{\kappa}$. These parameters are analogous to the inverse of dispersion measures in
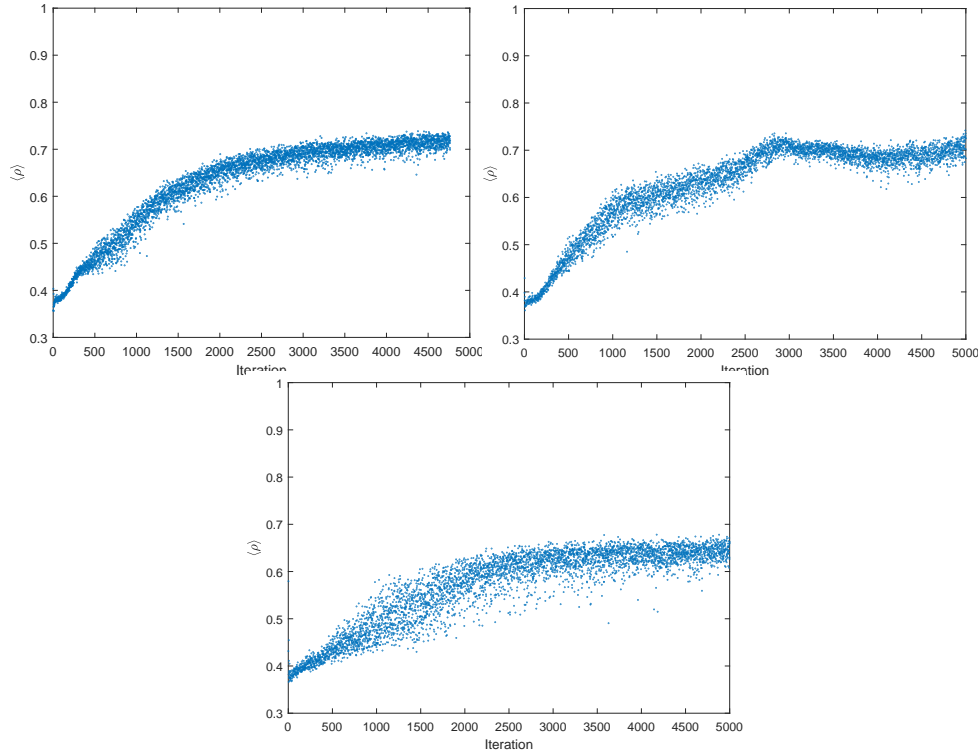
Figure 16: Evolution of average density in three runs initialized with the same seed. In each run, all the hyperparameters were fixed except the number of samples used in the Monte–Carlo estimates. (top left) 600, (top right) 400 and (bottom) 200 realization from the extended multivariate von Mises distribution are used.

descriptive statistics. As the algorithm progresses, the concentration of the distribution gradually increases, except κ_{ω_c}, representing the angle of rotation of the octagon, and stagnates close to zero. This behaviour can also be observed in the 1D histogram and 2D projections of 600 realization of the output distribution involving ω_c variable in Figure 3.

The evolution of the interaction matrix $\mathbf{D}$ is shown in Figure 15 for $\cos-\cos$ interactions, $\sin-\sin$ interactions and $\cos-\sin$ interactions. From the geometric interpretation of the exponential family selection quantile-based entropic trust region in subsection 3.4 perspective, we can see that the extended multivariate von Mises distribution is indeed moving towards a higher degree of interaction between angles. In the sense of unsupervised learning, the algorithm is learning a representation of the optimization landscape, and Figure 3 can be regarded as a visualization of this representation in the setting of $p2$-packings of regular octagons.

Appendix F. Influence of sampling pool on stability. The Fisher metric tensor $\mathcal{I}_{\boldsymbol{\theta}}$ used in the trust region step (3.7) has in the case of $p2$ group rank $r = 72$. Based on the ratio $c = 0.12$ of the Fisher matrix rank to the number of samples mentioned in Appendix D.6, the current implementation of the algorithm uses 600 samples generated at each iteration for the step size estimates in the densest $p2$-packing problem. Though less than unity, c cannot be considered a negligible quantity and results in problems with the variance of the sample covariance matrix estimator
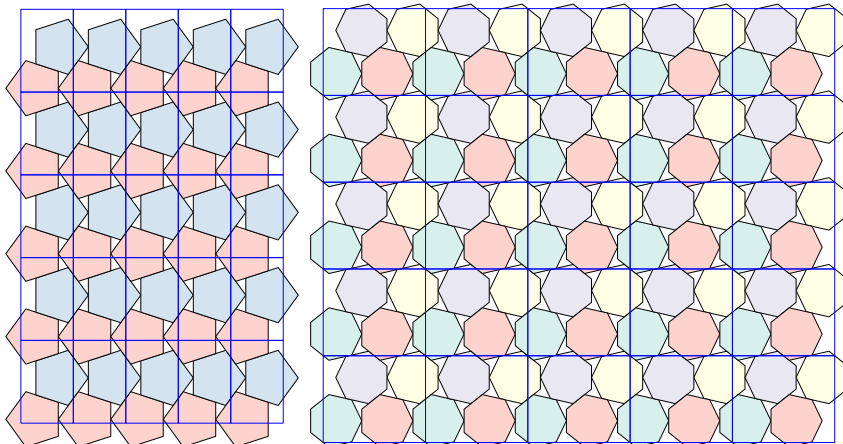
Figure 17: Visualization of 25 cells of the output configurations of the densest (left) regular pentagon *pg*-packing and (right) regular heptagon *p2gg*-packing.

of the Fisher matrix (A.1), even though we use adaptive learning rates and the Fisher metric scaling discussed in Appendix D.4.2 and Appendix D.5 to mitigate the effects of large estimator deviations.

To illustrate this issue, we compare three optimization runs with the random number generator initialized with the same seed and hyperparameter settings but a different number of samples generated at each iteration. Figure 16 shows evolution of the average density $\langle \rho \rangle$ where 600, 400 and 200 are used in the trust region step size Monte–Carlo estimates (A.3). Visually, the run where 200 samples are used is more scattered than in the cases with 400 and 600, which implies larger fluctuations in the trust region updates between iterations. Numerically, the variance of average density first differences $\Delta^t \langle \rho \rangle = \langle \rho \rangle^t - \langle \rho \rangle^{t-1}$ is 3.87 10^{-4} in the case of 600 samples, 4.5 10^{-4} for the 400 samples, and 1.2 10^{-3} for the 200 samples, and this supports our observations. The density of the best solution found in the case with the 600 samples is $\rho(\mathcal{K}_{p2}) = 0.897526117081202$, in the case of 400 samples $\rho(\mathcal{K}_{p2}) = 0.893840811317134$, and $\rho(\mathcal{K}_{p2}) = 0.890457479411185$ for the 200 samples. Although the variance of the average mean differences between the 600 and 400 sample instances is rather small, the highest density solution found during 5,000 iterations was in the case of the former.

Appendix G. Additional plane group packings. To demonstrate the robustness of the entropic trust region search, we present additional densest plane group experiments on a few selected convex polygons for which their densest packings are known. In all test cases, the hyperparameters are set according to Table 2. Numerical comparisons of the results are presented in Table 1.

G.1. *pg*-packing of regular pentagon. General optimal packing of congruent copies of the regular pentagon is believed to be a double lattice configuration with the density

$$\rho_{\mathrm{opt}} = \frac{5 - \sqrt{5}}{3} \approx 0.9213106$$

[36]. In terms of plane groups, a double lattice is the group *p2* already introduced in the *p2*-octagon packing case study in section 4. The *p2* group can be viewed as
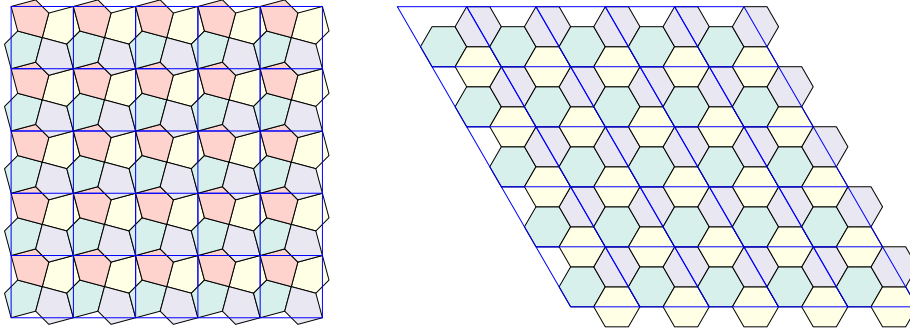
Figure 18: Visualization of 25 cells of the output configurations of the densest (left) irregular pentagon $p4$-packing and (right) regular hexagon $p3$-packing.

a collection of 2 lattices related to each other by a 2-fold rotational symmetry that permutes the two lattices.

Interestingly, our packing algorithm converged to the general optimal configuration when applied to search for the densest pg packings of regular pentagons. The pg plane group is not a semi-direct product of a lattice group and a point group since it contains a glide reflection symmetry operation that is neither an element of the point group nor the lattice group associated with pg. The crystal system is rectangular, meaning the angle ω_p between the primitive cell's basic vectors is fixed to $90°$, which reduces the number of optimization variables to 5 compared to the $p2$ plane group.

The output configuration of the densest pg-packing of the regular pentagon optimization schedule is visualized in Figure 17 with packing density $\rho(\mathcal{K}_{pg}) = 0.9213 - 1060131385$ and the theoretical optimum difference $\Delta_{\mathcal{K}_{pg}} = 7.2852\ 10^{-8}$, implying that the densest packing of regular pentagon can be realized using a glide reflection symmetry instead of the 2-fold rotational symmetry in $p2$ plane group.

G.2. $p2gg$-packing of regular heptagon. The optimal double lattice packing that is a $p2$-packing of the regular heptagon is

$$\rho_{\mathrm{opt}} = \frac{2}{97}\left(-111 + 492\cos\left(\frac{\pi}{7}\right) - 356\cos^2\left(\frac{\pi}{7}\right)\right) \approx 0.8926906$$

[42] and is conjectured to be the general optimal packing of regular heptagons.

The output configuration of our densest $p2gg$-packing optimization schedule is shown in Figure 17 with packing density $\rho(\mathcal{K}_{p2g}) = 0.89269066997639$ and theoretical optimum difference $\Delta_{\mathcal{K}_{p2gg}} = 1.6150\ 10^{-08}$, showing that the densest configuration of regular heptagons has higher symmetry than it was previously known.

$p2gg$ plane group symmetry operations are given by a 2-fold rotational symmetry and 2 glide reflection along perpendicular mirror planes. The crystal system is the same as in the pg group, that is rectangular, restricting the pg-packing configuration space dimension to 5 and the fractional coordinates of the heptagon to $0 \leq c_1 \leq \frac{1}{2}$ and $0 \leq c_2 \leq \frac{1}{2}$.

G.3. $p4$-packing of irregular pentagon. An irregular pentagon with the sequence of internal angles $120°$, $120°$, $90°$, $120°$, $90°$ tiles the euclidean plane, and its tiling is called the Cairo pentagonal tiling. Although the full symmetry of the Cairo tiling is $p4gm$, it is not a $p4gm$-packing but rather a $p4$-packing. Similarly to the $p2$ group, $p4$ is a semi-direct product of a point group with a 4-fold rotational symmetry and a lattice group belonging to the square crystal system, imposing restrictions on
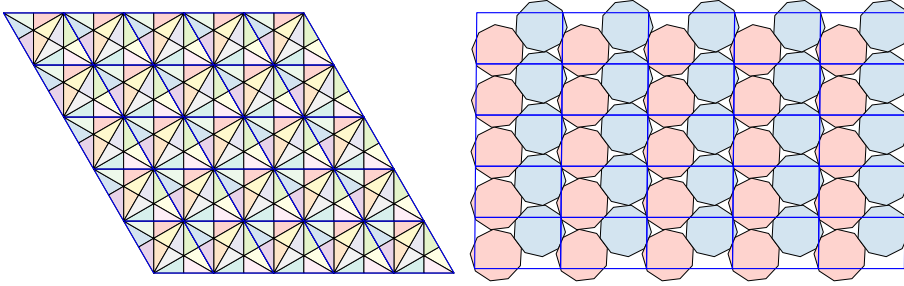
Figure 19: Visualization of 25 cells of the output configurations of the densest (left) $30 - 60 - 90$ triangle $p6mm$-packing and (right) regular enneagon $p2$-packing.

the shape of the primitive cell. That is, the sizes of the lattice group generators $\mathbf{b}_1$ and $\mathbf{b}_2$ are equal, and the angle between $\mathbf{b}_1$ and $\mathbf{b}_2$ fixed to $\omega_p = \frac{\pi}{2}$. Figure 18 shows the output configuration of our densest packing optimization procedure with density $\rho\left(\mathcal{K}_{p4}\right) = 0.99999999503997$ and theoretical optimum difference $\Delta_{\mathcal{K}_{p4}} = 4.9600\ 10^{-9}$.

G.4. $p3$**-packing of regular hexagon.** The regular hexagon tiles the 2D euclidean space, which corresponds to an optimal packing density of $\rho_{\mathrm{opt}} = 1$.

Figure 18 shows the output configuration when we applied the entropic trust region packing algorithm to the regular hexagon in the plane group $p3$. $p3$ is a semi-direct product of a lattice group and a point group consisting of a 3-fold rotational symmetry. The crystal system is hexagonal, which means that some restrictions on the shape of the primitive cell are imposed. Specifically, $\parallel \mathbf{b}_1 \parallel = \parallel \mathbf{b}_2 \parallel$, that is, the lattice group generators $\mathbf{b}_1$ and $\mathbf{b}_2$ are of equal size, and the angle between $\mathbf{b}_1$ and $\mathbf{b}_2$ is fixed to $120°$, reducing the degrees of freedom of the packing problem to 4. That is the fractional coordinates $0 \leq c_1 \leq \frac{2}{3}$ and $0 \leq c_2 \leq \frac{2}{3}$ of the triangle's centroid in the asymmetric unit, the angle ω_c of the rotation of the triangle, and the size of one of the edges of the primitive cell $\parallel \mathbf{b}_1 \parallel$.

The $p3$ group also forces additional linear constraints on the position of the triangle's centroid coordinates c_1 and c_2, which are

$$c_2 - \min\{1 - c_1, \frac{c_1}{2} + \frac{1}{2}\} \leq 0,$$
$$c_1 - \frac{c_2}{2} - \frac{1}{2} \leq 0.$$

These restrictions are treated as additional inequality constraints in (D.1) and are incorporated within the penalty function (D.2) computations.

The output packing density of our optimization schedule is $\rho\left(\mathcal{K}_{p1}\right) = 0.99999993-380570$ with the optimal packing difference of $\Delta_{\mathcal{K}_{p1}} = 6.6194\ 10^{-08}$. In practice, this configuration is a tiling showcasing some symmetries of regular hexagonal tiling.

G.5. $p6mm$**-packing of** $30 - 60 - 90$ **triangle.** The $30 - 60 - 90$ triangle is an irregular polygon with internal angles of $30°$, $60°$ and $90°$. Since it tiles the euclidean plane, its exact packing density is $\rho_{\mathrm{opt}} = 1$. The $p6mm$ group consist of a 6-fold rotational symmetry and reflections through 6 mirror planes. The crystal system is the same as in the $p3$ plane group mentioned earlier, which is hexagonal. The $p6mm$ group induces additional linear constraints on the position of the triangle's centroid

fractional coordinates c_1 and c_2 by

$$2c_1 - c_2 - 1 \leq 0,$$
$$-\frac{c_1}{2} + c_2 \leq 0.$$

and are treated as additional inequality constraints in (D.1) as in the case of $p3$-packing of hexagons Appendix G.4.

The output configuration of our optimization schedule applied to the densest $p6mm$-packing for the $30 - 60 - 90$ triangle search is shown in Figure 19 with packing density of $\rho(\mathcal{K}_{p6mm}) = 0.99999999871467$ and optimal packing difference $\Delta_{\mathcal{K}_{p6mm}} = 1.2853e - 09$, confirming that the $30 - 60 - 90$ triangle constitutes the primitive cell of the $p6mm$ plane group.

G.6. $p2$**-packing of regular enneagon.** Additionally, we applied the entropic trust region to the densest $p2$-packing of regular enneagon. Although theoretical optimal packing is unknown, the densest packing of regular enneagons with the $p2$ plane group symmetry, introduced in section section 4, was reported 0.901030078420934 in [24]. The density of the output configuration shown Figure 19 is $\rho(\mathcal{K}_{p2}) = 0.901030017272363$.