Greedy Geometric Algorithms for Collections of Balls, with Applications to Geometric Approximation and Molecular Coarse-Graining

F Cazals *and T. Dreyfus †and S. Sachdeva ‡and N. Shah §

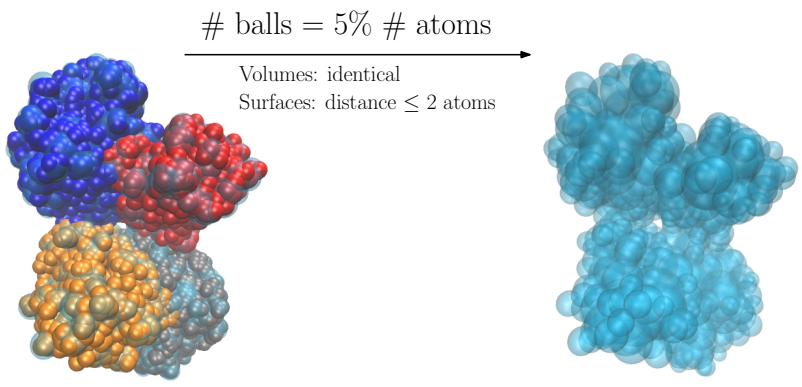
November 6, 2013

Abstract

Choosing balls which best approximate a 3D object is a non trivial problem. To answer it, we first address the *inner approximation* problem, which consists of approximating an object $\mathcal{F}_{\mathcal{O}}$ defined by a union of n balls with $k < n$ balls defining a region $\mathcal{F}_{\mathcal{S}} \subset \mathcal{F}_{\mathcal{O}}$. This solution is further used to construct an *outer approximation* enclosing the initial shape, and an *interpolated approximation* sandwiched between the inner and outer approximations.

The inner approximation problem is reduced to a geometric generalization of weighted max k-cover, solved with the greedy strategy which achieves the classical $1 - 1/e$ lower bound. The outer approximation is reduced to exploiting the partition of the boundary of $\mathcal{F}_{\mathcal{O}}$ by the Apollonius Voronoi diagram of the balls defining the inner approximation.

Implementation-wise, we present robust software incorporating the calculation of the exact Delaunay triangulation of points with degree two algebraic coordinates, of the exact medial axis of a union of balls, and of a certified estimate of the volume of a union of balls. Application-wise, we exhibit accurate coarse-grain molecular models using a number of balls 20 times smaller than the number of atoms, a key requirement to simulate crowded cellular environments.



Atomic model: 9,060 atoms

Interpolated shape: 450 balls

*Inria Sophia-Antipolis; Correspondance: frederic.cazals@inria.fr
†Inria Sophia-Antipolis
‡Princeton University
§Carnegie Mellon University

Contents

1 Introduction

1.1 Modeling with Balls

Three approximation problems.

Modeling complex 3D shapes is commonplace in science and engineering, and simple primitives such as balls play a central role in this process, for two reasons. On the one hand, the medial axis transform (MAT) allows representing a shape as a collection of balls [Ser82], usually infinite, so that sub-sampling such balls naturally yield approximations. On the other hand, (hierarchical) models represented by balls are ubiquitous, for example in molecular modeling, but also in robotics, computer graphics and CAGD, where bounding sphere hierarchies provide an elegant way to perform fast and numerically reliable collision detection. In this context, this paper addresses the following problems, which aim at approximating a given input shape by default (problem 1) and by excess (problem 2), and also finding an approximation sandwiched in-between the first two with a volume constraint (problem 3):

Problem. 1 Inner approximation *Given a 3D model $\mathcal{F}_\mathcal{O}$ consisting of the union of n balls, find a domain $\mathcal{F}_\mathcal{S} \subset \mathcal{F}_\mathcal{O}$, defined by the union of $k < n$ balls, such that the volume of $\mathcal{F}_\mathcal{O} \backslash \mathcal{F}_\mathcal{S}$ is minimized.*

Problem. 2 (Concentric) Outer approximation *Given a 3D model $\mathcal{F}_\mathcal{O}$ consisting of the union of n balls, find a domain $\mathcal{F}_\mathcal{S} \supset \mathcal{F}_\mathcal{O}$, defined by the union of $k < n$ balls, derived from an inner approximation. The approximation is termed* concentric *if the balls used to define $\mathcal{F}_\mathcal{S}$ are concentric with those of an inner approximation.*

Problem. 3 (Concentric) Interpolated approximation *Given a 3D model $\mathcal{F}_\mathcal{O}$ consisting of the union of n balls, find a domain $\mathcal{F}_\mathcal{S}$ sandwiched between an inner approximation and the associated outer approximation. The approximation is called* volume preserving *provided that $V(\mathcal{F}_\mathcal{S}) = V(\mathcal{F}_\mathcal{O})$. It is termed* concentric *if the balls used to define $\mathcal{F}_\mathcal{S}$ are concentric with those of an inner approximation.*

While we provide a general (and optimal) solution to problem 1, we only address the design of concentric outer and interpolated approximations. The reason for doing so is twofold. First, defining an outer approximation by growing the balls of an inner approximation defines a so-called Toleranced Model (TOM), namely a one-parameter family of shapes obtained by linearly interpolating the radii of the balls between the inner and outer radii. Moreover, a TOM is tantamount to a so-called additively-multiplicatively weighted Voronoi diagram, whose α-shape has also been characterized [CD10]. Thus, our algorithms allow studying a 1-parameter family of geometric approximations, rather than a single approximation, which is of importance whenever the objects studied are plagued with uncertainties, as in CAGD [LWC97] or structural biology [DDC12, DDC13]. Second and intuitively, growing the balls of an optimal inner approximation is an appealing strategy to build an outer approximation.

Since there is no ambiguity and for the sake of conciseness, the adjective concentric is omitted in the sequel.

Previous work. The approximation problems are actually connected to a variety of research veins, namely (i) geometric approximation algorithms for 3D shapes, (ii) robust geometric software development, (iii) approximation algorithms in general and weighted max k-cover in particular, and (iv) structural biology. We now briefly comment on recent work in these directions.

Geometric approximation algorithms for 3D shapes. In a broad perspective, the question of *sandwiching* a complex shape between an inner and an outer one is a classical problem in computer design where maximum and minimum material parts have been used in metrology (quality check) and robotics (collision detection) [LWC97]. While the particular class of shape used to define such approximations depends on the objects modeled, the medial axis transform (MAT) plays a fundamental role in defining a shape as a union of balls.

The particular case of a shape bounded by a smooth surface motivated the introduction of the MAT approximation using medial balls centered on specific Voronoi vertices called *poles* [AK00], an idea later re-used to approximate a shape bounded by a triangulated surface [AAK+09, SKS12]. This MAT approximation was also used for the *sphere-tree* construction [BO04], a representation to perform hierarchical

object modeling and collision detection, and to improve the grasping quality in robotics [PAD10]. For a shape with smooth boundary, the previous MAT approximation typically comes with a guarantee, namely the convergence of the Hausdorff distance between the input boundary and that of the approximation. Alternative methods skipping MAT computations have also been proposed. Of noticeable interest is [WZS+06], where an outer approximation for a model bounded by a triangle mesh is built, by combining sphere fitting and a greedy strategy to minimize the *sphere outside triangle volume* —see also section 5.

In a broader context, the problem of approximating a bounded open set has also been investigated recently. In [GMPW09] the authors introduce the scale-axis transform, which consists of scaling forth and back medial balls, so as to simplify a shape representation.

Robust geometric software development. It is worth noticing many of the works just mentioned rely on Voronoi diagrams, generally for the Euclidean distance, but also for a multiplicative distance in [GMPW09]. Consequently and from an implementation perspective, geometric algorithms from the Computational Geometry Algorithms Library [cga] (CGAL), but also number types from the `LEDA` [MN99] and `CORE` [KLPY99] libraries play a key role.

Approximation algorithms. The inner approximation problem is also related to approximation algorithms in general, and greedy strategies in particular. As we shall see, of particular interest is weighted max k-cover, which cannot be approximated within a ratio of $1 - 1/e + \varepsilon$ unless **P=NP** [Fei98], and the optimal bound $1 - 1/e$ is achieved by greedy strategies.

Structural biology. Last but not least, our incentive to address approximation problems for balls comes from computational structural biology, whose ultimate goal is to unravel the relationship between the structure and the function of macro-molecules. Originating with the work of Richards [LR71], molecular models represented as collections of van der Waals (vdW) balls and associated affine Voronoi diagrams have been instrumental to describe atomic packing properties [MJLC87, MLJ+87], to compute and decorate molecular surfaces [Con83, AE96], to exhibit correlations between structural and biological - biophysical properties of protein interfaces [BCRJ03, MDBC12], to select diverse conformational ensembles for mean field theory based docking algorithms [LSB+11], or to find entrance / exit passages to active sites [YFW+08].

While the aforementioned works are concerned with atomic resolution models, coarse-grain models are getting increasingly important to model isolated proteins or protein assemblies when partial or no atomic information is available, or when atomic models are too heavy to handle [Vak13].

More specifically, our incentive in developing accurate geometric approximations of molecules are related to two problems. The first one is the simulation of whole cellular environments [ME10, Goo09], using molecular dynamics or related techniques. (See also the beautiful illustrations of D. Goodsell at `http://mgl.scripps.edu/people/goodsell/books/MoL2-preview.html`.) These simulations require coarse-grain models since atomic resolution models of the individual molecules result in overly large models. However, because the dielectric coefficient of the water bulk is circa 40 times higher than that of the interior of a protein, these coarse-grain models must respect the atomic molecular volume as much as possible [Toz05] to retain accurate electrostatic interactions, a key component of the force field. (The dielectric coefficient is the screening term in Coulomb's equation.) More generally, the key features of a coarse-grain model, e.g. specified from its energy landscape (local minima corresponding to stable structures, and transitions between them) should match those of the associated atomic model [Cle07]. The second one deals with the modeling of macro-molecular machines involving from tens to hundreds of molecules. Modeling such machines relies on a panoply of complementary experimental techniques [ADV+07], resulting on noisy models due to a variety of uncertainties on the input data. To handle such models, in a spirit analogous to the maximum and minimum material parts used in mechanical engineering [LWC97], we introduced toleranced models based on balls (TOM) and established their correspondence to compoundly-weighted Voronoi diagrams [CD10]. In a nutshell, a TOM consists of a collection of pairs of concentric balls, where the inner balls cover a region of *high confidence* while the outer balls cover a volume bounding the model. We used TOM based on canonical proteins shapes (18 balls arranged on a lattice) to sharpen statistical analysis carried out on large assemblies [DDC12, DDC13]. Going beyond these results requires designing TOM of arbitrary geometry and topology, which in turn requires finding an inner and an outer approximation of a domain, using a collection of pairs of concentric balls. These are precisely the problems addressed in this paper.

1.2 Contributions

The key contributions of this work are to provide a provably correct optimal solution to the inner approximation problem, and to elaborate on this solution to design outer and interpolated approximations – although our outer approximation does not come with any theoretical guarantee. We also use volume preserving interpolated approximations to coarse-grain molecular models.

As we shall see, the design of an outer cover from an inner cover is conceptually simple, and merely relies on the Apollonius Voronoi diagram of the balls selected. Likewise, the design of an interpolated approximation merely requires a binary search to find the radii of balls interpolated between those of the inner and the outer approximations. However, the solution to the inner approximation problem is more complex, and actually relies on three contributions. First, we present a reduction of inner cover to a geometric generalization of weighted max k-cover involving a collection of balls related to the medial axis transform of the domain $\mathcal{F}_\mathcal{O}$. Second, we solve this geometric weighted max k-cover with the usual greedy strategy, showing that the $1 - 1/e$ classical bound known in combinatorial optimization applies, and also provide a lower bound on the volume of the selection $\mathcal{F}_\mathcal{S}$ w.r.t. the volume of $\mathcal{F}_\mathcal{O}$. From a combinatorial standpoint, our proofs are simplified versions of the classical ones for greedy algorithms and weighted max k-cover [NWF78, Propositions 4.1 and 4.3]. Yet, we include them for two reasons: first, it helps understanding the condition on the weights used in weighted max k-cover (their positivity is mandatory); second, in section 3.3, we re-use the skeleton of these proofs to characterize the result of the greedy strategy for inner covering, w.r.t the total volume of $\mathcal{F}_\mathcal{O}$ instead of the optimum. From a geometric approximation perspective, our results depart from previous work since we focus on an approximation guarantee obtained with a finite set of balls rather than asymptotically. Third, we present a robust and effective implementation of the greedy algorithm, incorporating the calculation of the exact Delaunay triangulation of points whose coordinates are degree two algebraic numbers, of the exact medial axis of a union of balls, and of a certified estimate of the volume of a union of balls.

1.3 Notations

A sphere and a ball are respectively denoted S_i and B_i. If $\mathcal{X}$ refers to a collection of balls, $\mathcal{F}_\mathcal{X}$ refers to the corresponding domain i.e. $\mathcal{F}_\mathcal{X} = \cup_{B_i \in \mathcal{X}} B_i$. Given a finite set E, the set of all subsets of E with k elements is denoted $\binom{E}{s}$. The volume of a 3D domain is denoted $\mathrm{Vol}(D)$.

In the sequel, we consider a domain $\mathcal{F}_\mathcal{O}$ defined as the union of the n balls of a set $\mathcal{O}$, and we wish to find a set of k balls $\mathcal{S}$ whose union $\mathcal{F}_\mathcal{S}$ defines an inner approximation of $\mathcal{F}_\mathcal{O}$. (As we shall see, in general, $\mathcal{S} \not\subset \mathcal{O}$.)

2 Algorithms: Design

2.1 Inner Approximation

Inner approximation and the MAT. The inner approximation problem is a natural geometric approximation problem, due to its connection to the medial axis of the domain $\mathcal{F}_\mathcal{O}$: the medial axis being the locci of centers of maximal balls associated with the domain $\mathcal{F}_\mathcal{O}$ [Ser82, AK01, CG06], any ball in $\mathcal{S}$ must be centered on the medial axis — any other ball is contained in a maximal ball. In the following, we sketch the two main steps of our solution to the inner covering problem.

Step 1: Defining a finite covering of $\mathcal{F}_\mathcal{O}$ based on its MAT. The boundary $\partial\mathcal{F}_\mathcal{O}$ of the domain $\mathcal{F}_\mathcal{O}$ consists of *spherical polygons* (2-cells), delimited by *circles* or *circle arcs* (1-cells), the latter ones being bounded by *boundary points* (0-cells). Generically, a boundary point is defined by the intersection of three spheres from $\mathcal{O}$. In our case, as proved in [AK01] and illustrated on Fig. 1, the MA consists of so-called singular simplices of the α-complex of $\mathcal{O}$ for $\alpha = 0$, together with the subset of the Voronoi diagram of the boundary points located within regular components of the α-shape. In particular, one and two dimensional faces of the MA define an infinite set of medial balls. Therefore, defining $\mathcal{S}$ from the MAT of $\mathcal{F}_\mathcal{O}$ is not straightforward since there is an infinite collection of balls to choose from. To resolve this difficulty, we prove that there exists a finite set of balls $\mathcal{C}$ associated with the MAT and defining a

covering of $\mathcal{F}_\mathcal{O}$. That is, we exhibit a collection of balls $\mathcal{C}$ such that $\mathcal{F}_\mathcal{O} = \mathcal{F}_\mathcal{C}$ (Lemma 2), so that $\mathcal{S}$ shall be a subset of $\mathcal{C}$ (Fig. 2).

Step 2: Solving the inner approximation using geometric weighted max k-cover. Consider the volumetric decomposition of $\mathcal{F}_\mathcal{O}$ induced by the spheres $S_i \in \mathcal{C}$. This decomposition is defined by the 3D arrangement of the spheres in $\mathcal{C}$: it consists of 3D cells $\mathcal{A} = \{A_i\}_{i=1,\ldots,m}$ induced by the spheres in $\mathcal{C}$, each cell A_i being contained in selected balls from $\mathcal{C}$. Also assume that we are given a weight function w, i.e. a real valued function defined over the cells of $\mathcal{A}$. Consider now the following maximization problem:

Problem. 4 *Given a weight function w, find a subset $\hat{\mathcal{S}}$ of $\mathcal{C}$ of size k, called the* selection*, such that:*

$$\hat{\mathcal{S}} = \arg\max_{\mathcal{S}\in\binom{\mathcal{C}}{k}} w(\mathcal{S}), \text{ with } w(\mathcal{S}) = \sum_{A_i \subset \mathcal{F}_\mathcal{S}} w(A_i). \tag{1}$$

Solving this problem when the weight function is the plain Euclidean volume of a cell of the arrangement $\mathcal{A}$ provides an inner approximation. In particular, we shall provide guarantees on the greedy solution w.r.t. the optimum, based on the analysis of weighted max k-cover (see below). We shall also extend these guarantees to compare the volume of the greedy solution against the volume of the input domain $\mathcal{O}$.

Complexity issues and the greedy strategy. Given an alphabet $\mathcal{A}$ of m points, and a collection $\mathcal{C}$ of subsets of $\mathcal{A}$, max k-cover aims at selecting k subsets from $\mathcal{C}$ so as to maximize the number of points from $\mathcal{A}$ which are covered [GJ79, Fei98]. (In the literature, this problem is sometimes called set cover [FG89]. To avoid confusion, we consider that the set cover problem aims at minimizing the number of sets in $\mathcal{C}$ to cover at least k elements from $\mathcal{A}$.) We note that the classical max k-cover is a special case of problem 4 with function w assigning a unit weight to all cells. Since weighted max k-cover is a **NP** complete problem, a polynomial time solution both in $|\mathcal{O}|$ and k cannot be expected. However, the problem is in **P** for a fixed k since all subsets of size k can be probed. But this brute force method is doomed to fail even for moderate k, which calls for alternate strategies, the greedy strategy being the most natural one.

The greedy strategy consists of k iterations, the jth step consisting of selecting the B_j maximizing the weight of the union of the balls selected so far. Because the selection obtained upon halting with k balls may not realize the optimum solution, the performance assessment of greedy relies on the worst-case ratio between the solution returned and the optimal one. For weighted max k-cover, this ratio is known to be of $1 - 1/e$, and is tight [CFN77, NWF78, FG89, Fei98].

Practically, a priority queue of the non-selected balls is maintained along the iterative selection process. The priority of a ball is the volume increment this ball would provide if selected, so that computing this priority only requires a function returning the volume of a union of balls [CKL11].

2.2 Outer Approximation

We derive our outer approximation from an inner one. To see how, recall that the Apollonius diagram of a collection of balls $\{B_i(c_i, r_i)\}$ is the Voronoi diagram defined for the following generalized distance [BWY06]:

$$\delta_i(p) = || \ p - c_i \ || - r_i. \tag{2}$$

Note that the Apollonius distance is merely the Euclidean distance from point p to the sphere S_i bounding B_i. For each ball B_i of the selection $\mathcal{S}$, consider the restriction of the boundary of the input domain not covered yet by the domain $\mathcal{F}_\mathcal{S}$, to its Voronoi cell $\mathrm{Vor}_{\mathrm{Apo.}}(B_i)$ in the Apollonius diagram. If this restriction is non empty, we define the expansion radius r_i^+ of S_i by the maximum distance of a point of that region to S_i, that is:

$$r_i^+ = \max_{p\in\partial\mathcal{F}_\mathcal{O}\cap\mathrm{Vor}_{\mathrm{Apo.}}(B_i) \text{ and } p\notin\mathcal{F}_\mathcal{S}} \delta_i(p). \tag{3}$$

If this restriction is empty, the expansion radius is set to the original radius, that is $r_i^+ = r_i$. Expanding each ball of the selection by its expansion radius yields an outer cover of the input domain (Fig. 3).

Practically, the expansion radii are computed via a discretization as a point cloud of the boundary $\partial\mathcal{F}_\mathcal{S}$ of the selected domain $\mathcal{F}_\mathcal{S}$ (and likewise for the input domain $\mathcal{F}_\mathcal{O}$), the Hausdorff distance between this

point cloud and the boundary surface being upper bounded by a parameter $\varepsilon_M \geq 0$. See the supplemental section 9.4 for details.

2.3 Interpolated Approximation

Consider an inner approximation together with the associated outer approximation as defined in section 2.2, and denote r_i and r_i^+ the inner and outer radii of the ith ball. Given a parameter $t \in [0, 1]$, we define the interpolated radius of the ith ball as

$$r_i(t) = (1 - t)r_i + tr_i^+. \tag{4}$$

An *interpolated approximation* is the union of these interpolated balls; it is called *volume preserving* if its volume matches that of the input shape.

2.4 Heuristic: Connecting an Approximation

If the input domain $\mathcal{F}_\mathcal{O}$ is connected, so should be the domain $\mathcal{F}_\mathcal{S}$: for example, the selection associated to a connected molecule should also be connected. To meet this constraint, the following heuristic may be used.

Let $\mathcal{S}_k$ be the selection upon termination, and consider the exposed balls i.e. the balls contributing to the boundary $\partial\mathcal{F}_{\mathcal{S}_k}$. Split these balls into two groups L and L^c, namely the largest component (in number of exposed balls), and the remaining ones. We aim at connecting L to one of the connected components of L^c. Consider the *intersection graph* with one vertex per ball $B_i \in \mathcal{C}$ and one edge for every pair of intersecting balls. Using this graph, we compute the shortest path joining any vertex representing a ball in L to any vertex representing a ball in L^c. This shortest path uses vertices representing balls in $\mathcal{C}\backslash\mathcal{S}_k$, which are added to the selection. This process is iterated until one connected component remains.

3 Inner Approximation: Guarantees

3.1 Defining the set $\mathcal{C}$ of step 1

As discussed when introducing problem 1, the inner approximation problem requires using balls centered on the medial axis of the domain $\mathcal{O}$, denoted MA for short in the sequel. But the medial axis is a cell complex with two dimensional faces, so that one has an infinite collection of balls to choose from. To circumvent this difficulty, consider the following classical lemma, related to pencils of spheres [Ber87]:

Lemma. 1 *Consider two intersecting spheres Σ_1 and Σ_2 in 3D, and define their convex linear combination, namely $\Sigma_\lambda = \lambda\Sigma_1 + (1 - \lambda)\Sigma_2$, with $\lambda \in [0, 1]$. The ball bounded by Σ_λ is contained in the union of the balls bounded by Σ_1 and Σ_2.*

Denote B_p^* a maximal ball centered on a vertex p of the medial axis, and let $\mathcal{C}$ be the set of all such balls. By the structure theorem of the medial axis of a union of balls [AK01], this set is finite. We shall use this set to run the greedy algorithm, since, as established by the following lemma, the balls in $\mathcal{C}$ define a covering of the input domain:

Lemma. 2 *The input domain $\mathcal{F}_\mathcal{O}$ satisfies*

$$\mathcal{F}_\mathcal{O} = \mathcal{F}_\mathcal{C}, \text{ with } \mathcal{C} = \{B_v^*\}_{v \text{ vertex of the MA of } \mathcal{F}_\mathcal{O}}. \tag{5}$$

Proof. [W]e shall prove that any maximal ball B_p^* is contained in the union of at most three balls centered on vertices from $\mathcal{C}$. Omitting the trivial case of a singular vertex of the medial axis, we first note that there are three cases to be analyzed, namely when p belongs to a singular edge of the medial axis, when it belongs to a (possibly clipped) Voronoi face f, or when it belongs to a singular triangle.

Case 1. This is exactly the case covered by lemma 1. In this case, the portion of the pencil contains the intersection circle between the two spheres defining the singular edge.

Case 2. The second case contains two sub-cases, namely when p lies in the interior of a Voronoi edge, and when p lies in the interior of the Voronoi facet f. The first sub-case is again the case of lemma 1 — all the spheres in the portion of the pencil contain the three boundary points defining the Delaunay triangle dual of the Voronoi edge in question. For the second one: let c be any Voronoi vertex of f belonging to $\mathcal{C}$, let L be the ray emanating from c and passing through p, and let d be the intersection point between L and the boundary ∂f of f. Point d belongs to either a Voronoi edge or to an α-shape edge (if the Voronoi facet is a clipped Voronoi facet in the medial axis). Call e and f the endpoints of this 1-cell of the medial axis. Now, by lemma 1, one has $B_d^* \subset B_e^* \cup B_f^*$ and similarly $B_p^* \subset B_c^* \cup B_d^*$. Thus, $B_p^* \subset B_c^* \cup B_e^* \cup B_f^*$.

Case 3. Amenable to the analysis carried out for Case 2.

Thus, since any maximal ball is contained in the union of at most three balls centered at vertices from $\mathcal{C}$, the claim holds. □

3.2 Greedy Strategy: Worst-case Bound w.r.t the Optimum

We now consider problem 4 with the following setting: the cells $\{A_i\}$ are those of the 3D arrangement induced by the balls in $\mathcal{C}$ (lemma 2); the weight function is some non-negative function. (Again, for inner cover, the plain Euclidean volume.)

To solve this problem with the greedy strategy, we use the following notation. The ball selected at the k^{th} step is denoted C_k, and the weight of the optimum set of balls OPT. Also, let $w^*(C_k)$ be the sum of the weights of the new elements covered by C_k that have not been covered in $C_j, 1 \leq j < k$ (i.e. the weight increment at step k). We start with a lemma (proof in appendix) needed to prove theorem 1.

Lemma. 3 *For $1 \leq i \leq k$, the following holds:*

$$w^*(C_i) + \frac{1}{k} \sum_{j=1}^{i-1} w^*(C_j) \geq \frac{OPT}{k}. \tag{6}$$

Using lemma 3, one proves (proof in appendix) :

Theorem. 1 *Consider the volumetric arrangement associated with a collection of balls $\mathcal{C}$, whose cells are equipped with non-negative weights. For Problem 4, the greedy approach has an approximation ratio of $1 - (1 - 1/k)^k > 1 - 1/e$.*

Moreover, the bound of Theorem 1 is tight: while this fact is a consequence of the hardness results [Fei98, Proposition 5.2], our proof is accompanied by an example achieving the lower bound (Fig. 8) (proof in appendix) :

Theorem. 2 *The greedy approach cannot perform better than $1 - (1 - 1/k)^k$.*

Remark. 1 *The proof of lemma 3 uses union-bound so that non-negativity assumption on the weights is mandatory. As a counter-example, consider the sets $C_1 = \{e1, e2\}, C_2 = \{e2, e3\}$ with $w(e1) = w(e3) = 1$ and $w(e2) = -1$. The union-bound fails for $w(C_1 \cup C_2)$. This remark is of particular interest in bio-physics, where atoms are decorated with physical, chemical or biological properties. For example, a weighting function that would take into account the electrostatic properties, which may be negative, would preclude the application of the previous lemma.*

3.3 Greedy Strategy: Worst-case Bound w.r.t the Total Weight

Approximation bound. The previous result can be generalized with respect to the weight of the whole input domain rather than that of OPT. We state the result for the particular case of the volume: (proof in appendix) :

Lemma. 4 *The volume of the selection $Vol(\mathcal{F}_{\mathcal{S}})$ and that of the input domain $Vol(\mathcal{F}_{\mathcal{O}})$ satisfy:*

$$\frac{Vol(\mathcal{F}_{\mathcal{S}})}{Vol(\mathcal{F}_{\mathcal{O}})} \geq 1 - \left(1 - \frac{1}{n}\right)^k \tag{7}$$

Tight example. Consider n disjoint balls of same radii. Then the greedy algorithm would select any k balls out of it. This would contribute a volume equal to k/n times the total volume. Also note that

$$\frac{k}{n} = 1 - \left(1 - \frac{k}{n}\right) \approx 1 - \left(1 - \frac{1}{n}\right)^k$$

for large values of n. This is optimal since no algorithm can approximate the union of n balls with approximation factor greater than k/n in this example, and thus in the worst case.

4 Algorithms: Implementation

We now provide an overview of the three approximation algorithms. The main steps undertaken, see Algorithm 1, follow the work-flow of section 2, as one successively deals with the construction of:

- The geometric structures underlying the three approximations, namely
  - The Delaunay triangulation DTB of the input balls, and the associated α-shape.
  - The Delaunay triangulation DTV of the boundary points of $\partial\mathcal{F}_{\mathcal{O}}$, and the dual Voronoi diagram DTV^*.
  - The medial-axis of the union of input balls.
  - The set $\mathcal{C}$ of candidate balls.
- The inner approximation.
- The outer approximation.
- The interpolated approximation.

The reader is also referred to the supplemental section 9 for a detailed description of the C++ classes involved in conjunction with the CGAL library [cga], in particular regarding the numerics.

5 Results

5.1 Dataset

Molecular models. As test set, we used the 96 protein - protein complexes from [LCJ99], available from the Protein Data Bank `http://www.rcsb.org/`. The complexes are of high biological interest since all of them are coupled to well identified biological processes. The number of atoms lies in the range $[1008, 13214]$, with a median of 3757. By default, a molecular model is defined as a so-called van der Walls (vdW) model, with atomic radii in the range $[1\mathring{A}, 2\mathring{A}]$. The so-called solvent accessible (SAS) model consists of expanding the atomic radii of the vdW model by the quantity $e = r_w = 1.4\mathring{A}$. This process mimicks a continuous layer of water molecules on the atoms, and allows recovering connexions between atoms nearby in 3D space, yet, not connected by covalent bonds. We also explore more general models using a radius expansion of $e \geq r_w$.

Molecular models are challenging both from the geometric and topological standpoint. Geometrically, side-chains of amino-acids sticking out of a protein are equivalent to the fingers of a character or the tail of an animal — that is molecules exhibit thin parts. Topologically, the Betti numbers $(\beta_0, \beta_1, \beta_2)$ of the models are usually large, witnessing many tunnels and cavities. The typical number of tunnels (β_1) and cavities (β_2) of a SAS model is of several tens, the molecular model of PDB file 1dhk.pdb being extreme, with 11 tunnels and 78 cavities. Notice in particular that tunnels and cavities are obstacles preventing using large balls to define an inner approximation.

Performance Enhancement via Dilation. Intuitively, the ability of the greedy algorithm to provide a good inner approximation relies on the possibility to choose large balls, which depends on two factors. First, the topological complexity: the closest to a topological ball the domain $\mathcal{F}_{\mathcal{O}}$, the better — high

Algorithm 1 Computing the inner, outer, and interpolated approximations of a domain $\mathcal{F}_\mathcal{O}$ defined as a union of balls: overview.

{**Problem specification: parameters**}
{Input balls $\mathcal{O}$ defining the domain $\mathcal{F}_\mathcal{O}$}
{Selection size k}
{Boolean flag to enforce the connectivity of $\mathcal{F}_\mathcal{S}$}
{Meshing precision ε_M for $\partial\mathcal{F}_\mathcal{O}$ and $\partial\mathcal{F}_\mathcal{S}$}

{**Pre-processing**}
Compute:
– The Delaunay triangulation DTB of the input balls $\mathcal{O}$, and the associated 0-shape
– the vertices of the boundary $\partial\mathcal{F}_\mathcal{O}$ of $\mathcal{F}_\mathcal{O}$
– the Delaunay triangulation DTV of these vertices, and the dual Voronoi diagram DTV^*
– the medial axis of $\partial\mathcal{F}_\mathcal{O}$ using DTB and DTV^*, using the algorithm described in [AK01]
– the balls in $\mathcal{C}$ defining the covering of $\mathcal{F}_\mathcal{O}$, as specified by lemma 2

{**Inner approximation**}
– Select the set $\mathcal{S}$ consisting of k balls amidst $\mathcal{C}$, using the greedy algorithm from section 2.1

{**Optional: connectivity enforcement**}
– Add balls from $\mathcal{C}\backslash\mathcal{S}$ to enforce the connectivity of $\mathcal{F}_\mathcal{S}$, as explained in section 2.4

{**Outer approximation**}
– Mesh the domains $\partial\mathcal{F}_\mathcal{O}$ and $\partial\mathcal{F}_\mathcal{S}$ with precision ε_M, using the `Mesher_3` mesher from the CGAL library
– Compute the expansion radii of the balls in the selection $\mathcal{S}$, as specified in section 2.2

{**Interpolated approximation**}
– Compute the interpolated radii of the balls defining the interpolated approximation, as specified in section 2.3

Betti numbers make the problem harder. Second, the geometric complexity: the more convex the domain $\mathcal{F}_{\mathcal{O}}$, the better. Along this line, enlarging the input balls by the quantity e discussed above results in the domain $\mathcal{F}_{\mathcal{O}}^{e}$, whose topology can be simpler than that of $\mathcal{F}_{\mathcal{O}}$. That is, the dilation may trigger the destruction of small cavities and tunnels, a statement which will be illustrated in section 5.4.

Selection size. In the tests, the selection size k is generally expressed in percents with respect to the model size. For example, $k/n = 5\%$ means that for a molecule of n atoms, a selection of size $k = n/20$ was used. The typical values used are $k/n \in \{1\%, 2\%, 5\%, 10\%\}$.

5.2 Monitoring Hausdorff Distances

Our strategy being volume based, to further assess it, we propose to compute the (signed) one-sided Hausdorff distances between the boundaries of the input domain and of the selection, respectively.

Distances between boundaries: the Hausdorff signature. The inner approximation being driven by a volumetric criterion, we further analyze the output in terms of one-sided Hausdorff distance (denoted $d_H(\cdot,\cdot)$) between the boundaries $\partial\mathcal{F}_{\mathcal{O}}$ and $\partial\mathcal{F}_{\mathcal{S}}$. More precisely, we code the position (interior versus exterior) of a point p with respect to a compact domain $\mathcal{F}$ by signing the one-sided Hausdorff distance, that is:

$$s(p, \partial\mathcal{F}) = \begin{cases} -\min_{q\in\partial\mathcal{F}} d(p,q) \text{ if } p \in \mathcal{F}, \\ +\min_{q\in\partial\mathcal{F}} d(p,q) \text{ otherwise,} \end{cases} \tag{8}$$

from which the relative position of $\partial\mathcal{F}_{\mathcal{O}}$ and $\partial\mathcal{F}_{\mathcal{S}}$ is defined by the following *Hausdorff signature*:

$$S_H(\partial\mathcal{F}_{\mathcal{O}}, \partial\mathcal{F}_{\mathcal{S}}) = [\min_{p\in\partial\mathcal{F}_{\mathcal{S}}} s(p, \partial\mathcal{F}_{\mathcal{O}}),\ \max_{p\in\partial\mathcal{F}_{\mathcal{S}}} s(p, \partial\mathcal{F}_{\mathcal{O}});\ \min_{p\in\partial\mathcal{F}_{\mathcal{O}}} s(p, \partial\mathcal{F}_{\mathcal{S}}),\ \max_{p\in\partial\mathcal{F}_{\mathcal{O}}} s(p, \partial\mathcal{F}_{\mathcal{S}})] \tag{9}$$

Note that for the inner approximation, the first two terms must be non-positive, while the last two terms must be non-negative. Note also that the maximum of the absolute values of the four terms is the Hausdorff distances between the two boundaries.

5.3 Timing statistics

The calculation of an interpolated approximation is summarized by the following signature, whose entries are expressed in seconds:

$$(t_P, t_{In}, t_C, t_M, t_{Out}, t_{Int}), \tag{10}$$

with t_P: time devoted to all preliminary geometric constructions (DTB, DTV, DTV^*, medial axis , set $\mathcal{C}$); t_{In}: time to run the inner selection with algorithm `Greedy`; t_C: time to connect the inner selection; t_M: time to mesh the boundary $\partial\mathcal{F}_{\mathcal{O}}$; t_{Out}: time to compute the outer approximation; t_{Int}: time to compute the interpolated approximation.

Three facts emerge from Table 1. First, the time consumed by the preliminary constructions (t_P) is negligible with respect to that of the approximation algorithms. Second, building the inner cover (t_{In}) is significantly more expensive than inferring the outer and the interpolated approximations (t_{Out} and t_{Int}) from the inner cover. Third, the limiting step at this stage is the boundary meshing (t_M). This owes to the precision imposed on the mesh ($\varepsilon_M = 0.2$; atomic radii in the SAS model are in he range $[2.4\mathring{A}, 3.4\mathring{A}]$), and to the genericity of the mesher. (The algorithm meshes implicit surfaces in general, and is constrained to respect circle-arcs and vertices found on the boundary of the union.) A meshing algorithm dedicated to the boundary of the union would certainly yield an improvement of one or two orders of magnitude, but the focus of this work being on approximation guarantees, we leave this improvement for further work.

5.4 Approximations

Inner Approximation. In a vdW model, only balls of covalently bonded atoms intersect. Thus, for a vdW model, one expects the volume covered to vary linearly as a function of the selection size, which is exactly observed (Fig. 5, red curves). Consider now the SAS model of a given vdW model. We discussed

in section 5.1 the expected benefits associated to model dilation. This phenomenon is precisely observed, since the larger r, the larger the candidate balls, and the better the volume ratio curve (Fig. 5 again).

To assess the efficacy of inner approximations for SAS models, we observe that selections with $k/n \geq 5\%$ are such that the volume ratio between the inner cover and the input domain is always above 0.65, with a median equal to 0.77 (Fig. 4, red error bars). Using $k/n = 5\%$, it is also observed that the connectedness of the inner approximation is often verified by the output from `Greedy`, since the minimum, median and maximum number of balls added by the heuristic of section 2.4 are respectively 0, 1, and 5. Two inner approximations with a selection with $k/n = 5\%$ are illustrated on Figs. Fig. 6(B) and Fig. 7(B).

Outer approximation. The evolution of volume ratio upon increasing the selection size shows that for $k/n \geq 5\%$, the ratio is always below 1.61 (Fig. 4, green error bars). Two outer approximations with a selection with $k/n = 5\%$ are illustrated on Figs. Fig. 6(B) and Fig. 7(B).

We also compared our outer approximation scheme against that of [WZS$^+$06]. To this end, we picked one small (PDB id: 3sgb) and one large (PDB id: 1fin) protein complexes, dilated by $e = 5.6$ to eliminate small tunnels and cavities — topological features are not mentioned for the models used in [WZS$^+$06]. To compare our results against those of [WZS$^+$06, Fig. 17], Table 3 displays the *relative inside volume* E_R^-, namely the percentage of missing volume (fourth column of Table 3), and the *relative outside volume* E_R^+, namely the percentage of excess volume (fifth column of Table 3). While the models processed are quite different, our outer approximation compares favorably against that from [WZS$^+$06, Fig. 17], which we illustrate with the extreme selection sizes used in [WZS$^+$06]: for 16 balls, the statistic E_R^+ is in the range $[120\%, 400\%]$ for [WZS$^+$06], and less than 65% for us; for 128 balls, E_R^+ is in the range $[20\%, 40\%]$ for [WZS$^+$06], and less than 27% in our case. Furthermore, our running time is comparable to the ones reported in [WZS$^+$06, Fig. 9] (more than 400 seconds for 16 balls and more than 1400 for 128 balls).

Interpolated approximation and coarse-grain molecular models. Motivated by the structural biology applications discussed in introduction, we computed interpolated approximations for the 96 protein complexes, with expansion radii $e = r_w = 1.4\mathring{A}$ (SAS model), and $e = 5.6\mathring{A}$ (a value meant to study the performances on models with fewer tunnels and cavities). Selection size of 1%, 2%, 5% and 10% were used. Since the volume of the input model is conserved, our assessment is based on the four-tuple of Eq. (9), denoted (d_1, d_2, d_3, d_4) for the sake of conciseness.

Consider Tab. 2. For $e = 1.4$ and a selection size $\geq 2\%$, the Hausdorff distances correspond to less than two atoms line-up in the SAS model. The relatively large values of d_1 (resp. d_3) are accounted for by topological features (tunnels, cavities) of the interpolated approximation (resp. input model) inside yet not present in the input model (resp. interpolated approximation). For $e = 5.6$, all Hausdorff distances are less than the diameter of an atom in the SAS model, illustrating the fact that for simpler topologies (no tunnel and no cavity), the aforementioned difficulties do not arise.

We illustrate these results with the interpolated approximations of two systems at selection size $k/n = 5\%$. The first one is a globular protein of 1690 atoms (3sgb) whose SAS model contains three tunnels and five cavities (Fig. 6). Approximation-wise, the interpolated approximation made of 85 balls has no tunnel but one cavity. Its Hausdorff signature is $[-4.30, 2.31; -3.31, 1.63]$. The second one is a larger complex of 9060 atoms, whose SAS model contains 20 tunnels and 70 cavities (Fig. 7). Approximation-wise, the interpolated approximation consisting of 453 balls has 32 tunnels and 15 cavities. Its Hausdorff signature is $[-5.65, 5.12; -9.01, 3.42]$.

6 Conclusion and Outlook

This paper studies three basic geometric approximation problems for a collection of balls, namely inner and outer covering, as well as the problem of designing a volume preserving geometric approximation. The inner approximation problem is shown to be a geometric version of weighted max k-cover, defined on a collection of balls associated with the medial axis transform of the input domain, for which a greedy strategy can be used. The outer approximation problem reduces to computing the partition of the boundary of the original model by the Apollonius Voronoi diagram of the balls of the inner approximation. Finally, computing the volume preserving interpolated approximation reduces to finding an approximation sandwiched between the inner and outer approximations.

It is also shown that the best possible approximation factor for inner approximation $(1 - 1/e)$ is retained by the greedy strategy, a result which we extend for the output of greedy with respect to the total volume of the input domain. Our implementations hinge upon state-of-the-art software coupled to the CGAL library, as they involve the exact calculation of a Delaunay triangulation for points whose coordinates are degree two algebraic numbers, the intersection of the dual of this triangulation with the α-complex of the input balls, and the certified calculation of the volume of a union of medial balls. This implementation handles molecular models containing up to $O(10^5)$ atoms within minutes (inner approximation, and interpolated approximation given the inner and outer approximations). For these reasons, we believe that our algorithm should prove useful for a broad class of geometric approximation problems dealing with balls, in particular in the context of approximate medial axis transforms, where the focus has been so far on asymptotic properties—upon increasing the number of balls.

Yet, our work calls for further developments, both in the theoretical and applied directions. On the theoretical side, two challenging questions are of high interest. First, our greedy algorithm comes with guarantees for the inner approximation problem, a property stemming from the relationship between inner cover and the medial axis transform of the shape, which allows phrasing the problem as geometric weighted max k-cover. But coming up with other guarantees, namely bounding the excess volume of the outer cover, and controlling the volume of the symmetric difference between the input domain and that the selection (or controlling their Hausdorff distance) for the interpolated cover are open problems.

Second, constraining the geometric selection by topological criteria, e.g. prescribed Betti numbers, would also be of the highest interest. However, approximation problems aiming at accommodating both geometric and topological criteria are likely to be challenging—it has been shown that the so-called homology localization problem is **NP**-hard. In an applied vein and as mentioned in introduction, we believe that a key application of our algorithms will be the design of coarse-grain macro-molecular models, to investigate macro-molecular machines and simulate crowded environments within whole cells. But prior to undertaking these challenges, one will have to decorate our purely geometric coarse-grain models with bio-physical properties, while retaining the essential properties of the corresponding atomic models.

Acknowledgments. Michael Hemmer is acknowledged for his help with the CGAL kernels, and the reviewers are acknowledged for insightful comments. This work has partially been supported by the EC STREP project CGL, (contract No. 255827).

References

[AAK+09] O. Aichholzer, F. Aurenhammer, B. Kornberger, S. Plantinga, G. Rote, A. Sturm, and G. Vegter. Recovering structure from r-sampled objects. In *Computer Graphics Forum*, volume 28, pages 1349–1360. John Wiley & Sons, 2009.

[ADV+07] F. Alber, S. Dokudovskaya, L. M. Veenhoff, W. Zhang, J. Kipper, D. Devos, A. Suprapto, O. Karni-Schmidt, R. Williams, B.T. Chait, M.P. Rout, and A. Sali. Determining the Architectures of Macromolecular Assemblies. *Nature*, 450(7170):683–694, Nov 2007.

[AE96] N. Akkiraju and H. Edelsbrunner. Triangulating the surface of a molecule. *Discrete Applied Mathematics*, 71(1-3):5–22, 1996.

[AK00] N. Amenta and R.K. Kolluri. Accurate and efficient unions of balls. In *Proceedings of the sixteenth annual symposium on Computational geometry*, pages 119–128. ACM, 2000.

[AK01] N. Amenta and R. K. Kolluri. The medial axis of a union of balls. *Comput. Geom. Theory Appl.*, 20:25–37, 2001.

[BCRJ03] R.P. Bahadur, P. Chakrabarti, F. Rodier, and J. Janin. Dissecting subunit interfaces in homodimeric proteins. *Proteins: Structure, Function, and Bioinformatics*, 53(3):708–719, 2003.

[Ber87] M. Berger. *Geometry I*, volume 1. Springer, 1987.

[BO04] G. Bradshaw and C. O'Sullivan. Adaptive medial-axis approximation for sphere-tree construction. *ACM Transactions on Graphics (TOG)*, 23(1):1–26, 2004.

[BWY06] J-D. Boissonnat, C. Wormser, and M. Yvinec. Curved voronoi diagrams. In J.-D. Boissonnat and M. Teillaud, editors, *Effective Computational Geometry for curves and surfaces*. Springer-Verlag, Mathematics and Visualization, 2006.

[CCLT09] P. M. M. De Castro, F. Cazals, S. Loriot, and M. Teillaud. Design of the cgal spherical kernel and application to arrangements of circles on a sphere. *Computational Geometry: Theory and Applications*, 42(6-7):536–550, 2009.

[CD10] F. Cazals and T. Dreyfus. Multi-scale geometric modeling of ambiguous shapes with toleranced balls and compoundly weighted α-shapes. In B. Levy and O. Sorkine, editors, *Symposium on Geometry Processing*, pages 1713–1722, Lyon, 2010. Also as INRIA Tech report 7306.

[CFN77] G. Cornuejols, M.L. Fisher, and G. Nemhauser. Location of bank accounts to optimize float: An analytic study of exact and approximate algorithms. *Management Science*, 23(8):789–810, 1977.

[CG06] F. Cazals and J. Giesen. Delaunay triangulation based surface reconstruction. In J.-D. Boissonnat and M. Teillaud, editors, *Effective Computational Geometry for curves and surfaces*. Springer-Verlag, Mathematics and Visualization, 2006.

[cga] CGAL, Computational Geometry Algorithms Library. http://www.cgal.org.

[CKL11] F. Cazals, H. Kanhere, and S. Loriot. Computing the volume of union of balls: a certified algorithm. *ACM Transactions on Mathematical Software*, 38(1):1–20, 2011.

[Cle07] C. Clementi. Coarse-grained models of protein folding: toy models or predictive tools? *Current Opinion in Structural Biology*, 17:1–6, 2007.

[Con83] M. L. Connolly. Analytical molecular surface calculation. *J. Appl. Crystallogr.*, 16(5):548–558, 1983.

[DDC12] T. Dreyfus, V. Doye, and F. Cazals. Assessing the reconstruction of macro-molecular assemblies with toleranced models. *Proteins: structure, function, and bioinformatics*, 80(9):2125–2136, 2012.

[DDC13] T. Dreyfus, V. Doye, and F. Cazals. Probing a continuum of macro-molecular assembly models with graph templates of sub-complexes. *Proteins: structure, function, and bioinformatics*, 2013. In press.

[Fei98] U. Feige. A threshold of ln n for approximating set cover. *Journal of the ACM*, 45(4):634–652, 1998.

[FG89] PC Fishburn and WV Gehrlein. Pick-and choose heuristics for partial set covering. *Discrete Applied Mathematics*, 22(2):119–132, 1989.

[GJ79] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, New York, NY, 1979.

[GMPW09] J. Giesen, B. Miklos, M. Pauly, and C. Wormser. The scale axis transform. In *ACM Symp. on Computational Geometry*, pages 106–115, 2009.

[Goo09] D. Goodsell. *The machinery of life*. Springer, 2009.

[KLPY99] V. Karamcheti, C. Li, I. Pechtchanski, and C. Yap. A core library for robust numeric and geometric computation. In *15th ACM Symp. on Computational Geometry, 1999*, pages 351–359, 1999.

[LCJ99] L. Lo Conte, C. Chothia, and J. Janin. The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology*, 285:2177–2198, 1999.

[LR71] B. Lee and F. M. Richards. The interpretation of protein structure: Estimation of static accessibility. *J. Molecular Biology*, 55:379–400, 1971.

[LSB+11] S. Loriot, S. Sachdeva, K. Bastard, C. Prevost, and F. Cazals. On the characterization and selection of diverse conformational ensembles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2):487–498, 2011.

[LWC97] J-C. Latombe, R. Wilson, and F. Cazals. Assembly sequencing with toleranced parts. *Computer Aided Design*, 29(2):159–174, 1997.

[MDBC12] N. Malod-Dognin, A. Bansal, and F. Cazals. Characterizing the morphology of protein binding patches. *Proteins: structure, function, and bioinformatics*, 80(12):2652–2665, 2012.

[ME10] S.R. McGuffee and H. Elcock. Diffusion, crowding and protein stability in a dynamic molecular model of the bacterial cytoplasm. *PLoS Comput. Biol.*, 6(3):1–18, 2010.

[MJLC87] S. Miller, J. Janin, A.M. Lesk, and C. Chothia. Interior and surface of monomeric proteins. *Journal of molecular biology*, 196(3):641–656, 1987.

[MLJ+87] S. Miller, A.M. Lesk, J. Janin, C. Chothia, et al. The accessible surface area and stability of oligomeric proteins. *Nature*, 328(6133):834–836, 1987.

[MN99] K. Mehlhorn and S. Näher. *LEDA: a platform for combinatorial and geometric computing*. Cambridge University Press, 1999.

[NWF78] G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

[PAD10] M. Przybylski, T. Asfour, and R. Dillmann. Unions of balls for shape approximation in robot grasping. In *Proc. IEEE Conf. Intelligent Robots and Systems (IROS)*, pages 1592–1599, 2010.

[Ser82] J. Serra. *Image Analysis and Mathematical Morphology*. Academic Press, London, UK, 1982.

[SKS12] S. Stolpner, P. Kry, and K. Siddiqi. Medial spheres for shape approximation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(6):1234–1240, 2012.

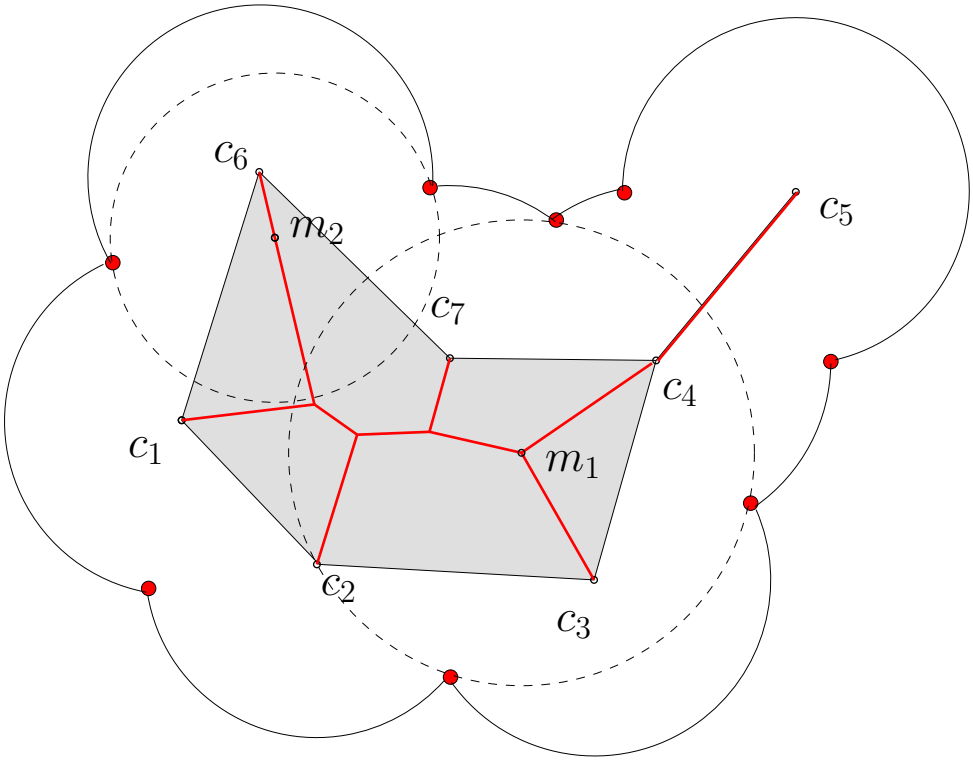
[Toz05] Valentina Tozzini. Coarse-grained models for proteins. *Current opinion in structural biology*, 15(2):144–150, 2005.

[Vak13] I. Vakser. Low-resolution structural modeling of protein interactome. *Current Opinon in Structural Biology*, 23:198–205, 2013.

[WZS+06] R. Wang, K. Zhou, J. Snyder, X. Liu, H. Bao, Q. Peng, and B. Guo. Variational sphere set approximation for solid objects. *The Visual Computer*, 22(9):612–621, 2006.

[YFW+08] E. Yaffe, D. Fishelovitch, HJ. Wolfson, D. Halperin, and R. Nussinov. Molaxis: Efficient and accurate identification of channels in macromolecules. *Proteins*, 73(1):72–86, 2008.

Figure 1 The medial axis transform for a union of balls: 2D illustration. The domain $\mathcal{F}_{\mathcal{O}}$ is defined by the union of 7 balls. Its boundary points are represented by red dots, while its medial axis (MA) is presented by red line-segments. Two maximal balls centered at m_1 and m_2 on the MA are presented in dashed circles.



.

Table 1 Minimum, median and maximum running times. The following parameters were used to process the database of 96 proteins complexes $\varepsilon_M = 0.2$, $k/n = 5\%$ and $e = r_w = 1.4$. The columns represent the entries Eq. (10).

Statistical	t_P	t_{In}	t_C	t_M	t_{Out}	t_{Int}
minimum	1.00	161.75	0.00	1499.17	6.21	2.89
median	2.10	392.10	40.78	3062.61	96.55	5.51
maximum	6.55	1151.75	917.57	9312.90	1513.78	15.82

Figure 2 The finite covering of the domain $\mathcal{F}_{\mathcal{O}}$ of Fig. 1 based on its medial axis transform. The set $\mathcal{C}$ consists of 7+4 balls: the seven blue balls are input balls — they contribute to the boundary $\partial\mathcal{F}_{\mathcal{O}}$; the four red balls are not input balls.

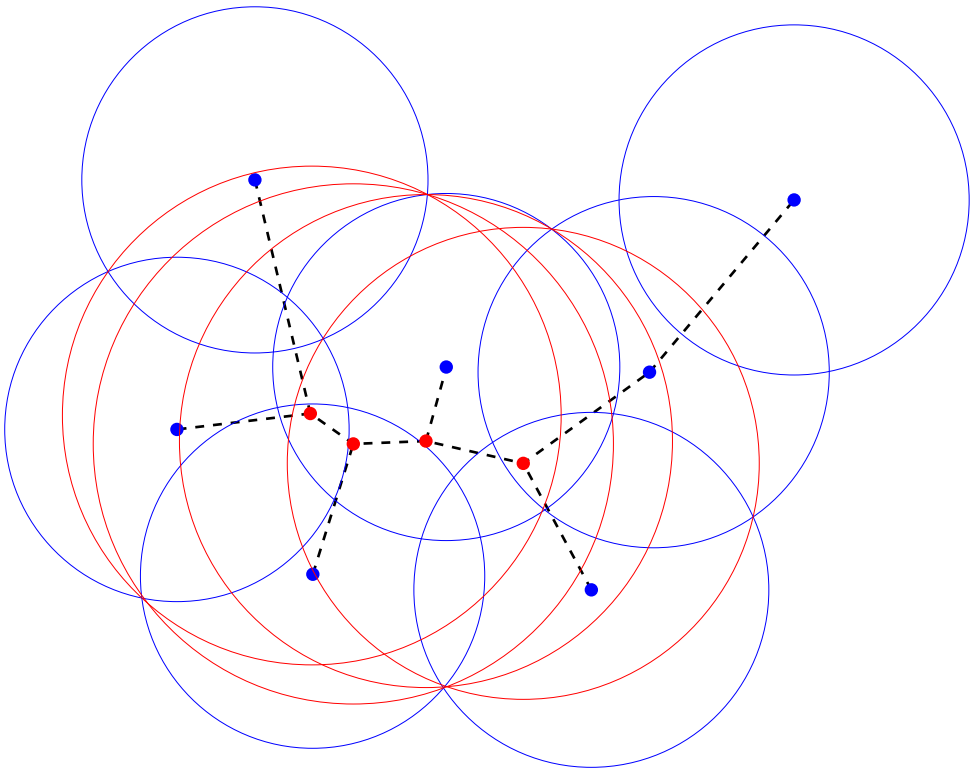


Figure 3 Computing an outer cover of the domain $\mathcal{F}_{\mathcal{O}}$ of Fig. 1, from an inner cover. Solid circles: the selection $\mathcal{S}$ defining the inner cover. Dashed circles: set $\mathcal{C}\backslash\mathcal{S}$. Black curves: Apollonius Voronoi diagram of $\mathcal{S}$. For each selected ball $S_i \in \mathcal{S}$, a point maximizing the Apollonius distance to S_i is shown in black, used to define the outer cover materialized by the arrows. Note that this point belongs to the Apollonius Voronoi cell of S_i.

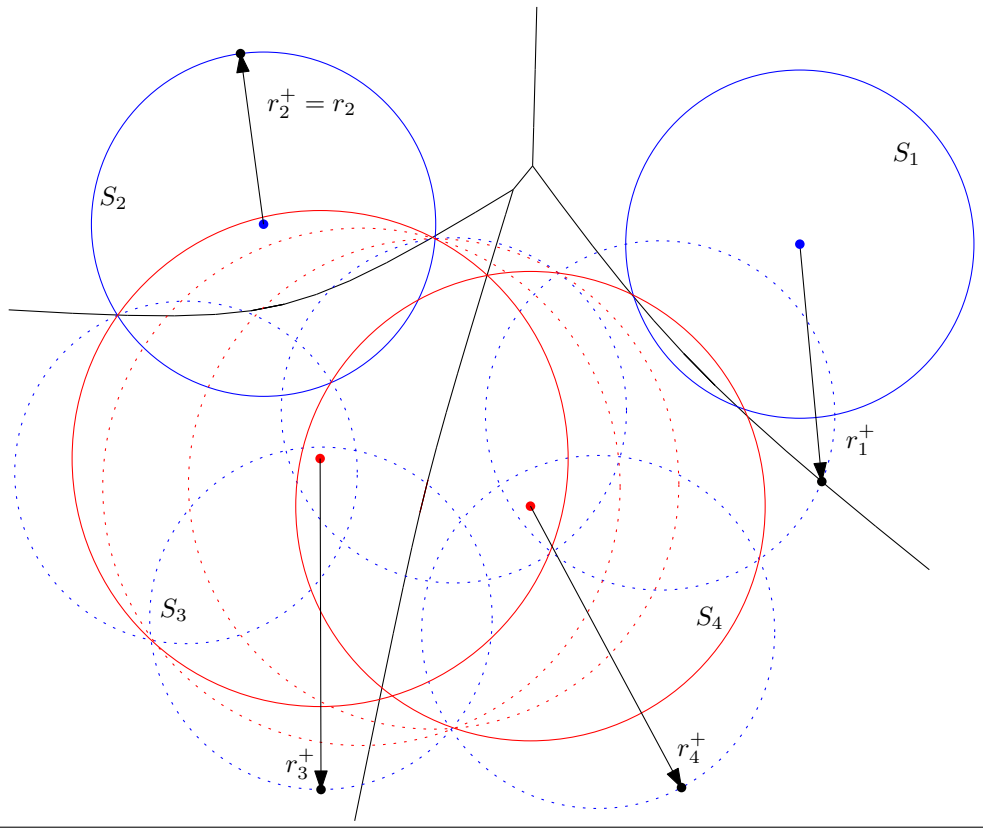


Figure 4 Inner and outer approximations: volume ratio w.r.t. the input shape, as a function of the selection size. The molecular models are the solvent accessible (SAS) ones.

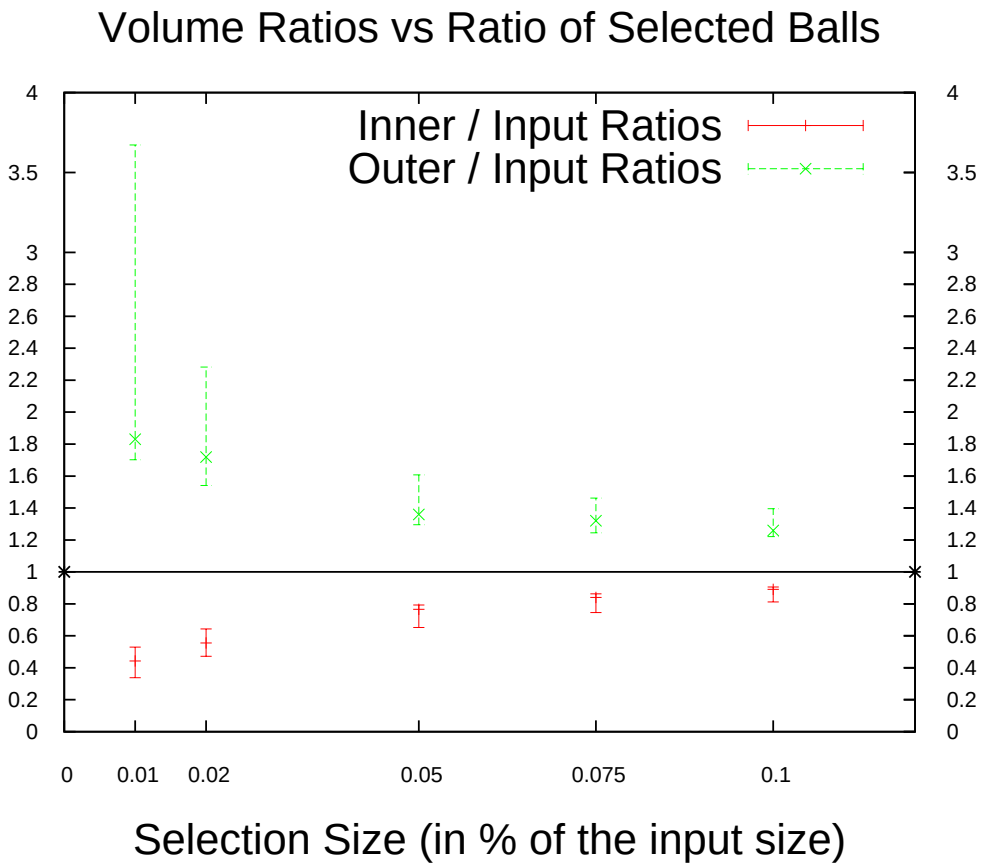
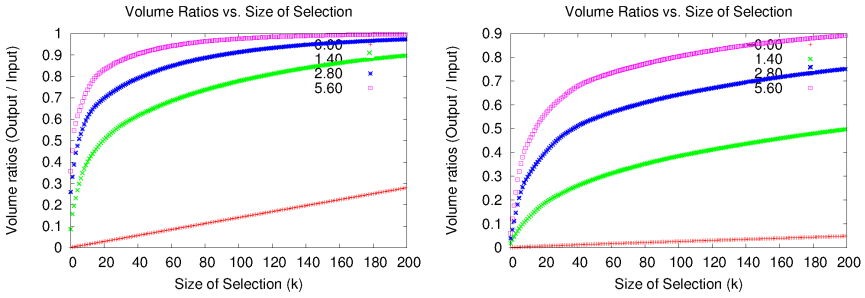


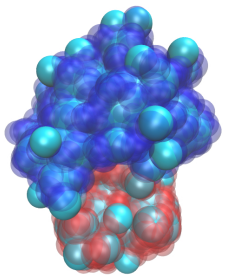
Figure 5 Inner approximation: volume ratio $\mathbf{Vol}(\mathcal{F}^e_{\mathcal{S}})/\mathbf{Vol}(\mathcal{F}^e_{\mathcal{O}})$ as a function of the expansion radius. (Left) Protein complex of 1690 balls (PDB code 3sgb, see Fig. 6). **(Right)** Protein complex of 9060 balls (PDB code 1fin, see Fig. 7)



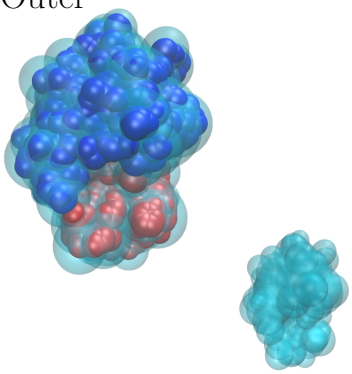
7 Artwork

Figure 6 Inner / Outer / Interpolated approximations with a selection size $k/n = 5\%$, for a small globular protein complex (PDB id: 3sgb). The atomic model contains 1,690 atoms, colored by their polypeptide chain. Each inset shows the approximation, the associated main figure displaying the superposition of the approximation and the atomic SAS model. (NB: the visual effect of inner balls sticking out from the model comes from the fact that some balls are common. The same holds for balls shared by the outer approximation and the model.)

(A) Inner



(B) Outer



(C) Interpolated

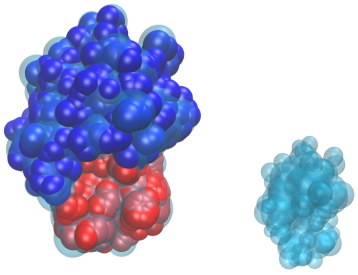


Figure 7 Inner / Outer / Interpolated approximations with a selection size $k/n = 5\%$, for a larger protein complex (PDB id: 1fin). The atomic model contains 9,060 atoms, colored by their polypeptide chain. Each inset shows the approximation, the associated main figure displaying the superposition of the approximation and the atomic SAS model.

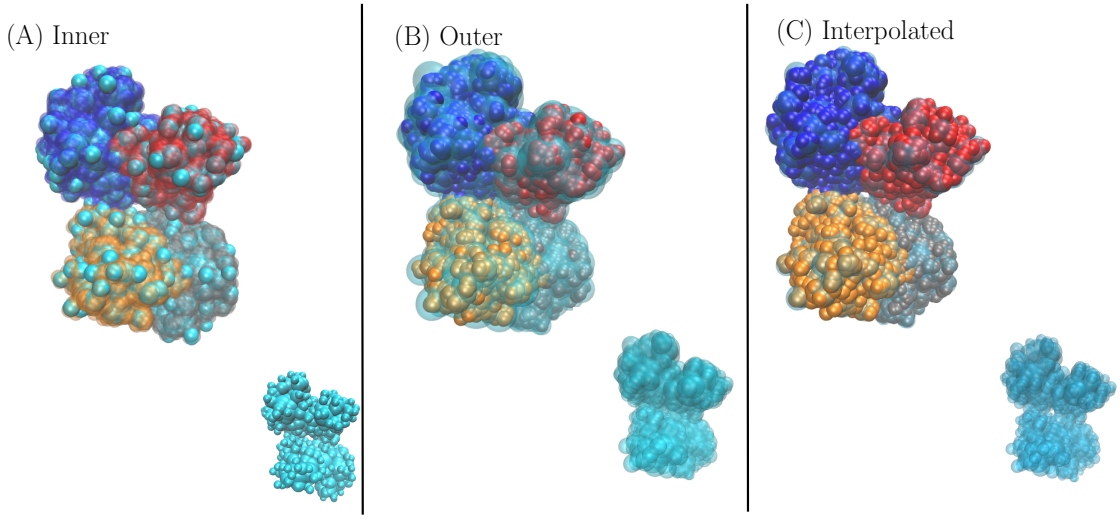


Table 2 Interpolated approximations: monitoring the signed one-sided Hausdorff distances as a function of the selection size k. Columns read as (1) Expansion radius e (2) Selection size k versus number of atoms n (3–6) The four terms of $S_H(\partial\mathcal{F}^r_{\mathcal{O}}, \partial\mathcal{F}^r_{\mathcal{S}})$ in Eq. (9), denoted d_1, d_2, d_3, d_4 for the sake of conciseness. Recall that d_1, d_2 characterize the boundary of the interpolated approximation w.r.t that of the input domain, and vice-versa for d_3, d_4.

e	k/n	d_1	d_2	d_3	d_4
r_w	0.01	-8.39 ± 1.76	7.26 ± 1.74	-6.12 ± 1.77	5.54 ± 1.38
r_w	0.02	-7.64 ± 1.76	5.46 ± 1.11	-7.11 ± 2.41	4.89 ± 1.63
r_w	0.05	-5.61 ± 1.63	2.94 ± 0.85	-7.43 ± 2.38	4.76 ± 2.44
r_w	0.10	-4.05 ± 1.71	2.77 ± 1.52	-7.80 ± 1.80	5.25 ± 2.23
r_w	mean	-6.48 ± 2.42	4.66 ± 2.30	-7.10 ± 2.21	5.11 ± 1.98
5.6	0.01	-3.17 ± 0.88	3.49 ± 0.34	-4.36 ± 0.78	2.43 ± 0.24
5.6	0.02	-2.25 ± 1.54	2.58 ± 0.22	-3.55 ± 0.61	1.49 ± 0.15
5.6	0.05	-0.91 ± 0.35	1.68 ± 0.14	-2.77 ± 1.11	0.65 ± 0.91
5.6	0.10	-0.38 ± 0.12	1.08 ± 0.13	-1.68 ± 0.47	0.28 ± 0.07
5.6	mean	-1.92 ± 1.44	2.41 ± 0.89	-3.33 ± 1.20	1.38 ± 0.94

Table 3 Statistics on the inner / outer / interpolated approximations of 3sgb and 1fin with $e = 5.6$ and $\varepsilon_M = 1$. Columns read as (1) PDB id of the protein (2) Selection size k versus number of atoms n (3) Betti numbers of the input model (4) Relative inside volume (percentage of missing volume) (5) Relative outside volume (percentage of excess volume) (6) running time signature – Eq. (10) (7) total running time.

PDB	k/n	$(\beta_0.\beta_1, \beta_2)$	E_R^-	E_R^+	$(t_P, t_{In}, t_C, t_M, t_{Out}, t_{Int})$	total time
3sgb	16/1690	$(1, 0, 0)$	20.06%	59.96%	$(0.43, 77.20, 0., 38.92, 0.02, 0.52)$	117.09
3sgb	128/1690	$(1, 0, 0)$	1.76%	22.30%	$(0.40, 1739.54, 0., 42.17, 0.97, 2.78)$	1785.86
1fin	16/9060	$(1, 1, 0)$	33.45%	64.46%	$(2.05, 106.03, 0., 190.42, 0.07, 1.58)$	300.05
1fin	128/9060	$(1, 1, 0)$	9.47%	26.99%	$(2.05, 1142.67, 0., 193.58, 3.43, 4.10)$	1345.83

8 Appendix to Section 2: Inner Approximations: Guarantees

8.1 Proof of lemma 3

Proof. [A]t the i^{th} step, we select C_i that maximizes the weight of the new cells A_j being covered. Because the balls selected up to step $i-1$ may cover cells which are not covered by the balls accounting for OPT, the weight of the cells that are covered by the optimum solution but not yet covered by the $(i-1)$ is at least

$$OPT - \sum_{l=1}^{i-1} w^*(C_l) \tag{11}$$

Since w is non-negative, the union-bound property states that for any collection of balls $C_1, \ldots, C_p$, one has $w(C_1 \cup \cdots \cup C_p) \leq \sum_{l=1,\ldots,p} w(C_l)$. Since all the cells involved in Eq. (11) are covered by the optimum set of balls, by the union-bound property, there must exist one ball, not yet selected, that covers these new cells with total weight at least

$$\frac{1}{k}\left(OPT - \sum_{l=1}^{i-1} w^*(C_l)\right). \tag{12}$$

Since C_i maximizes the weight of the new cells being covered, we must have

$$w^*(C_i) \geq \frac{1}{k}\left(OPT - \sum_{l=1}^{i-1} w^*(C_l)\right). \tag{13}$$

Rearranging completes the claim.
□

8.2 Proof of theorem 1

Using Lemma 3, the proof of Thm. 1 goes as follows:
Proof. [W]e show the following by induction:

$$\sum_{j=1}^{i} w^*(C_j) \geq \left(1 - \left(1 - \frac{1}{k}\right)^i\right) OPT \tag{14}$$

The property holds for $i = 1$ thanks to lemma 3.

Assuming that it holds at rank i, to see that it also holds at rank $i+1$, one multiplies Eq. (14) by $1 - 1/k$, and adds up the inequality obtained to that of lemma 3 for $i+1$.

For $i = k$, Equation (14) yields

$$\frac{\sum_{j=1}^{k} w^*(C_j)}{OPT} \geq \left(1 - \left(\frac{k-1}{k}\right)^k\right) \tag{15}$$

The left hand side is the ratio of the weight of the subset of $\mathcal{O}$ chosen by the greedy approach and the optimum solution i.e. that approximation factor and hence we have the above theorem. The fact that the above ratio is greater than $1 - \frac{1}{e}$ for all k is a trivial exercise. □

8.3 Proof of theorem 2

Theorem 2 and the following proof are illustrated by Fig. 8:
Proof. [F]ix a given k. We shall construct an example where the greedy approach can achieve an approximation ratio arbitrarily close to $1 - (1 - \frac{1}{k})^k$.

Let

$$\mathcal{A} = \{A_i\}_{i=1,\ldots,(k^2+k)}$$

$$\forall i,j \text{ s.t. } 0 \leq i < k, 1 \leq j \leq k,\ w(A_{i.k+j}) = \frac{1}{k^2}\left(\frac{k-1}{k}\right)^i$$

$$\forall j \text{ s.t. } 1 < j \leq k,\ w(A_{k^2+j}) = \frac{1}{k}\left(\frac{k-1}{k}\right)^k - \epsilon$$

The balls are defined as follows

$$\mathcal{O} = \{C_i\}_{i=1,\ldots,2k}$$

$$\forall i \text{ s.t. } 1 \leq i \leq k,\ C_i = \bigcup_{j=(i-1).k+1}^{i.k} A_j$$

$$\forall i \text{ s.t. } \leq i \leq k,\ C_{k+i} = \bigcup_{j \equiv i \pmod{k}} A_j$$

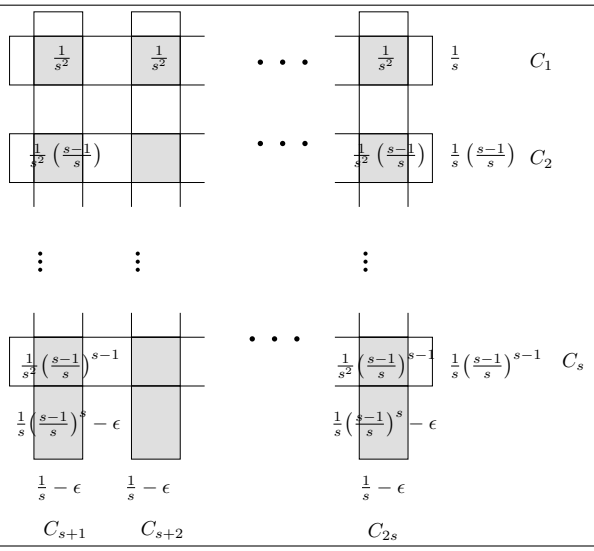
Simple calculations lead us the following total weights:

$$\forall 1 \leq i \leq k,\ w(C_i) = \frac{1}{k}\left(\frac{k-1}{k}\right)^{i-1}$$

$$\forall 1 \leq i \leq k,\ w(C_{k+i}) = \frac{1}{k} - \epsilon$$

The optimum choice of $\mathcal{S}$ with $|\mathcal{S}| = k$ is clearly $\{C_i\}_{i=k+1,\ldots,2k}$ with total weight $1 - k\epsilon$, whereas the greedy method would choose $\{C_i\}_{i=1,\ldots,k}$, with a maximum weight of $1-(1-\frac{1}{k})^k$, giving an approximation factor is arbitrarily close to $1 - (1 - \frac{1}{k})^k$. □

Figure 8 A tight example for the greedy strategy.



8.4 Proof of lemma 4

*Proof. [I]*n the proof of the approximation factor of the greedy algorithm for the volumetric decomposition given in lemma 3, note that it is valid for any solution and not only the optimum solution, i.e. no property of the optimum solution is required. Thus we replace the optimum solution by a solution selecting the given n balls. Thus we get the following equation.

$$w^*(C_i) \geq \frac{1}{n}\left(V - \sum_{j=1}^{i-1} w^*(C_j)\right)$$

where C_j is the j^{th} ball selected by the greedy algorithm, and $w^*(C_i)$ is the new volume of C_i not covered by any of C_j, $1 \leq j < i$. Solving it in the manner similar to that used in the proof of Thm 2 yields:

$$GREEDY = \sum_{i=1}^{k} w^*(C_i) \geq V \cdot \left(1 - \left(1 - \frac{1}{n}\right)^k\right)$$

□

9 Appendix to Section 4: Algorithms: Implementation

9.1 Inner approximation

Overview. The input consists of a collection of balls $\mathcal{O}$ defining a region $\mathcal{F}_\mathcal{O}$, and of a selection size k or a target ratio τ between the volume of $\mathcal{F}_\mathcal{S}$ and that of $\mathcal{F}_\mathcal{O}$ — that is one expects $\mathrm{Vol}(\mathcal{F}_\mathcal{S})/\mathrm{Vol}(\mathcal{F}_\mathcal{O}) \geq \tau$. The output consists of an ordered set of balls $\mathcal{S} \subset \mathcal{C}$, together with the increment in volume associated to each ball. (Recall that in general the set $\mathcal{C}$ is different from the set of input balls $\mathcal{O}$.)

The algorithm consists of iteratively selecting the ball providing the best volume increment, selected from a priority queue containing all candidates from the set $\mathcal{C}$, as specified in section 2.1. Upon selecting ball say B_i, we recompute the volume increments of all candidate balls intersecting B_i.

Note that as a preprocessing, we compute the *intersection graph*, namely the graph with one vertex per ball $B_i \in \mathcal{C}$, and one edge for every pair of intersecting balls. Incidences in this graph are used to identify the balls whose volume increments get recomputed upon selecting a particular ball.

We now present the geometric objects used by the algorithms, following the flow presented in section 2.1, and mentioning the CGAL (`http://www.cgal.org`) classes used and their template parameters when appropriate.

The Delaunay triangulation DTB of the input balls, and the associated α-shape. Following classical usage, we call K the kernel used to instantiate the CGAL classes `Regular_triangulation_3` and `Alpha_shape_3`. Two options for K are discussed below.

The Delaunay triangulation DTV of the boundary points of $\partial\mathcal{F}_\mathcal{O}$. We compute the MA of the input shape by restricting the Voronoi diagram DTV^* of the boundary points located within regular components of the α-shape, as recalled in section 2.1. The Voronoi diagram DTV^* is the dual of the Delaunay triangulation DTV of the boundary points of $\partial\mathcal{F}_\mathcal{O}$. Two difficulties are faced to construct DTV. First, more than three co-planar points are generic in DTV [AK01]. Second, since a boundary point is found at the intersection of three input spheres, its coordinates are degree two algebraic numbers. We therefore store these points using the CGAL spherical kernel `Spherical_kernel_3` [CCLT09], instantiated with K. The two options for K, referred to as the *inexact* and the *exact* kernels in the sequel, are:

- `Exact_predicates_inexact_constructions_kernel`, the underlying number type (NT) to store the coordinates of the boundary points being a `double`.
- `Exact_predicates_exact_constructions_kernel_with_sqrt`, the underlying number type to store the coordinates being either `CORE::Expr` or `LEDA::real`.

Additionally, a map is used to associate a singular or regular facet from the α-shape of DTB to each boundary point.

To handle these difficulties, we implemented a dedicated kernel denoted `DTV_kernel`, defining a new point type for the boundary points. This kernel is actually templated by two parameters:

- First, a ball identifier, used to record the three input spheres defining a boundary point. These identifiers are used to handle the aforementioned special cases, so as to avoid the numerical calculation of a predicate whose sign can be inferred from the fact that the input points lie on a set of known input spheres. Practically and since an input ball corresponds to a vertex of the α-shape of DTB, the vertex handle of the α-shape is taken as identifier.

- Second, a number type used to represent the coordinates of the boundary points, the two options being the NT associated to the aforementioned inexact and exact kernels.

One comment is in order about the Voronoi diagram DTV^*, which is the dual of DTV, since medial balls associated to selected Voronoi vertices are used by greedy. With the inexact kernel, the input points of DTV are approximations of the exact boundary points, since the degree two algebraic number get converted to doubles. For these points, the combinatorial structures of DTV and DTV^* are exact (exact predicates are used), but the embedding of the Voronoi vertices of DTV^* is inexact (inexact constructions are used). With the exact kernel, the input points of DTV are exactly the boundary points. Moreover, the embedding of the Voronoi vertices is exact (exact constructions are used).

The medial-axis of the union of input balls. We store the medial axis as a container of polygons, possibly degenerate for singular vertices and edges of the α-shape [AK01]. Our polygon class inherits from the CGAL class `Polygon_2` (embedded in 3D), instantiated with the kernel K. It offers new features, in particular the computation of the maximal ball centered at a point of the polygon. Such a ball has a center which is a `Point_3` from K, and a squared radius whose type is `NT`.

The set $\mathcal{C}$ of candidate balls. Following the results of section 3.1, the candidate balls used are only centered on the vertices of the medial axis. Such balls are associated with the medial axis, as just discussed.

The volume of the selected balls. Computing the volume of a union of balls is a difficult problem, from a combinatorial, but also numerical standpoint—inverse trigonometric functions are involved. We use our certified algorithm [CKL11] which returns an interval certified to contain the exact volume. More precisely, due to the impossibility to obtain a volume as an exact number type, whatever the kernel used (exact, inexact), the centers and radii of the candidate balls are converted to doubles. These balls are input to our algorithm, which requires two template parameters: the number type of the output (double or interval), and the level of exactness used to compute the constructions involved in the volume computation, namely the coordinates of Voronoi vertices, and boundary points of the union of the selected balls. Following the discussion in [CKL11], the three options are referred to as (faster, ck_pt_exact and all_exact). Practically, we use the pair (double, faster) for the inexact kernel, and (interval, all_exact) for the exact kernel.

9.2 Outer approximation

To compute the expansion radii of Eq. (3) without computing the partition of the boundary of the input object with respect to an Apollonius Voronoi diagram, we resort to discretization. Assume that $\partial\mathcal{F}_\mathcal{O}$ has been sampled, and denote $P_{\partial\mathcal{F}_\mathcal{O}}$ the corresponding point cloud. We assume that for some $\varepsilon_M > 0$, the one-sided Hausdorff distance between the boundary and the samples satisfies $d_H(\partial\mathcal{F}_\mathcal{O}, P_{\partial\mathcal{F}_\mathcal{O}}) \leq \varepsilon_M$. (See also section 9.4.) Let $C_p \subset P_{\partial\mathcal{F}_\mathcal{O}}$ be a point set initially consisting of the points from the sampled boundary not covered by $\mathcal{F}_\mathcal{S}$. Let C_B be the set of balls to be expanded, initialized as the subset of balls from the selection contributing to $\partial\mathcal{F}_\mathcal{S}$. For a given ball $B_i \in C_B$, we proceed in two stages. First, the point in $C_p \cap \mathrm{Vor}_{\mathrm{Apo.}}(B_i)$ maximizing $\delta_i(p)$ is computed. To account for the discretization, the corresponding additive distance is increased by ε_M. Second, the ball B_i is removed from C_B, and all points in $C_p \cap \mathrm{Vor}_{\mathrm{Apo.}}(B_i)$ are removed from C_p. The process is iterated until exhaustion of C_p.

Note that the previous algorithm does not require computing $\mathrm{Vor}_{\mathrm{Apo.}}(B_i)$, since the assignment of a point to its Apollonius Voronoi cell only requires computing its additive distances to all balls in C_B.

9.3 Interpolated approximation

Increasing the value of t in Eq. (4) yields nested balls, whence nested interpolated approximations. Therefore, finding the volume preserving interpolated approximation requires a binary search on $t \in [0, 1]$. Practically, the binary search is stopped when the discrepancy between the volumes is less than $\varepsilon_V = 10^{-5}$.

9.4 Effective Computation of the Hausdorff Distance and Expansion Radii

Hausdorff distances. To compute the terms of Eq. (9), assume that $\partial\mathcal{F}_\mathcal{O}$ and $\partial\mathcal{F}_\mathcal{S}$ have been sampled, and denote $P_{\partial\mathcal{F}_\mathcal{O}}$ and $P_{\partial\mathcal{F}_\mathcal{S}}$ the corresponding point clouds. We assume that for some $\varepsilon_M > 0$, one has:

$$d_H(\partial\mathcal{F}_\mathcal{O}, P_{\partial\mathcal{F}_\mathcal{O}}) \leq \varepsilon_M \text{ and } d_H(\partial\mathcal{F}_\mathcal{S}, P_{\partial\mathcal{F}_\mathcal{S}}) \leq \varepsilon_M. \tag{16}$$

Under the assumptions, two applications of the triangle inequality show that each term of the four-tuple of Eq. (9) is approximated in absolute value up to $2\varepsilon_M$. Practically, having sampled the boundaries using the CGAL mesher Mesher_3, computing an approximation of the signature of Eq. (9) requires two primitives, that is finding the nearest sample of a sample $p \in P_{\partial\mathcal{F}_\mathcal{O}}$ in $P_{\partial\mathcal{F}_\mathcal{S}}$ (and vice-versa), and checking whether p belong to the interior of a ball of the domain bounded by $\mathcal{F}_\mathcal{S}$. These primitives are easily implemented using the point location strategy of Delaunay_triangulation_3. Practically, the value $\varepsilon_M = 0.2$ was used.

9.5 Geometric Kernels: Performances and Robustness

Following the best practices in computational geometry, we designed a generic CGAL based implementation, and instantiated it with the aforementioned exact and inexact kernels. We compared the volume ratios obtained with these two kernels on a set of 10 protein complexes, and did not observe any difference before the third digit.

For running times, we compared the execution time for the construction of DTB, DTV and the medial axis. The selection itself was excluded from the timing, as also noticed in section 9.1, since our volume computation algorithm uses `double` as number type. On the aforementioned 10 models, we observed that the exact kernel was on average about 150 times slower than the inexact one. For these two reasons — absence of obvious degeneracies and much better running time, the results reported in the sequel were computed with the inexact kernel.

Using this inexact kernel, it is observed that the running times for computing DTB and DTV are a mere order of magnitude slower than the CGAL ones (`http://www.cgal.org/Manual/latest/doc_html/cgal_manual/Triangulation_3/Chapter_main.html#Subsection_39.6.1`) for the regular triangulation case (supplemental Fig. 9). These running times are naturally consistent with the fact that the geometric objects manipulated behave nicely for our molecular models: both the number of boundary points (supplemental Fig. 10) and the primitives of the medial axis (supplemental Fig. 11) are linear in the number of input balls.

10 Appendix to Section 7: Artwork

Figure 9 Running times for the key steps of the inner approximation algorithm, as a function of the number of input balls. The models processed are the Solvent Accessible Ones. (i) Delaunay triangulation DTB of input balls (ii) Delaunay triangulation DTV of boundary vertices (iii) Medial axis of the union of balls.

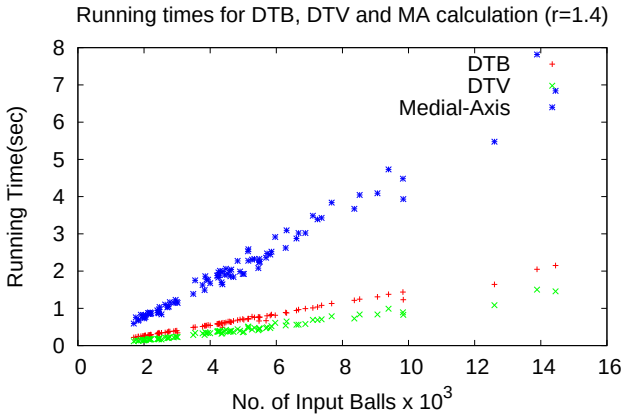


Figure 10 Number of boundary vertices as a function of the number of input balls. The models processed are the Solvent Accessible Ones.

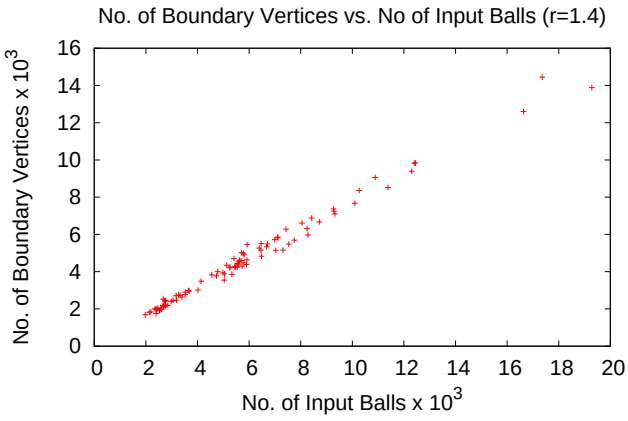


Figure 11 Number of faces of the medial axis as a function of the number of input balls. The models processed are the Solvent Accessible Ones.