

New error measures and methods for realizing protein graphs from distance data

CLAUDIA D'AMBROSIO¹, VU KHAC KY¹, CARLILE LAVOR², LEO LIBERTI¹, NELSON MACULAN³

¹ CNRS, UMR 7161 LIX, École Polytechnique, F-91128 Palaiseau CEDEX, France
Email:{dambrosio,vu,liberti}@lix.polytechnique.fr

² IMECC, University of Campinas, Brazil
Email:clavor@ime.unicamp.br

³ COPPE, Federal University of Rio de Janeiro, RJ 21941-972, Brazil
Email:maculan@cos.ufrj.br

June 15, 2016

Abstract

The interval Distance Geometry Problem (*i*DGP) consists in finding a realization in \mathbb{R}^K of a simple undirected graph $G = (V, E)$ with nonnegative intervals assigned to the edges in such a way that, for each edge, the Euclidean distance between the realization of the adjacent vertices is within the edge interval bounds. Our aim is to determine the best formulation-based method for this problem restricted to the application of finding protein conformation given some of the inter-atomic distances. To this end, we introduce (a) a new error measure, which could be described as a “root mean square deviation modulo isomers”; (b) a set of new and existing quadratic and semidefinite programming formulations of this problem; (c) a set of new and existing methods for solving these formulations. We then perform a computational evaluation of all the feasible solver+formulation combinations according to new and existing error measures, finding that the best methodology is a new heuristic method based on multiplicative weights updates.

Keywords: distance geometry, protein conformation, mathematical programming.

1 Introduction

The Distance Geometry Problem (DGP) is defined formally as follows: given an integer $K > 0$, a simple undirected graph $G = (V, E)$, and an edge weight function $U : E \rightarrow \mathbb{R}_+$, establish or deny the existence of a vertex realization function $x : V \rightarrow \mathbb{R}^K$ such that:

$$\forall \{u, v\} \in E \quad \|x_u - x_v\|_2 = U_{uv}; \quad (1)$$

realizations satisfying (1) are called *valid realizations*. The DGP arises in many important applications: determination of protein conformation from distance data [36], localization of mobile sensors in communication networks [16], synchronization of clocks from phase information [44], control of unmanned submarine fleets [6], spatial logic [17], and more [28]. It is **NP**-complete when $K = 1$ and **NP**-hard for larger values of K [43]. Notationwise, we let $n = |V|$ and $m = |E|$.

Let \tilde{X} be the set of valid realizations of G : then, if $x \in \tilde{X}$, obviously any congruence (translation, rotation, reflection) of x yields another valid realization of G . We therefore focus on the quotient set $X = \tilde{X}/\sim$, where $x \sim y$ whenever there is a congruence mapping x to y . Then $X = \emptyset$ if the corresponding DGP instance is a NO one, G is *rigid* if $|X|$ is finite, and *flexible* if $|X|$ is uncountable.

The only remaining possibility for $|X|$ is to be countably infinite, but this cannot happen. By Milnor's theorem on the Betti numbers of real algebraic varieties [38], the number of connected components of X

is bounded above by $2 \times 3^{nK-1}$. Suppose that $|X|$ is countably infinite: then it cannot be flexible. This implies that incongruent elements of X are on distinct connected components of the manifold containing X . Milnor’s theorem shows that there are only finitely many such connected components, which implies that $|X|$ is finite. This result also follows by the cylindrical decomposition theorem of semi-algebraic sets [8].

Our motivating application is protein conformation, where $K = 3$ and the given graph G belongs to a specific subclass of Henneberg type I graphs [45], namely there is an order $<$ on V such that, for each $v > 3$, v is adjacent to $v - 1, v - 2, v - 3$ [21]. DGP instances with this property form a problem called DISCRETIZABLE MOLECULAR DISTANCE GEOMETRY PROBLEM (DMDGP), which is also **NP**-hard [21]. In [27], we proposed a fast and accurate mixed-combinatorial algorithm for solving the DMDGP, called Branch-and-Prune (BP). Unsurprisingly, the BP has exponential complexity in the worst case, but the DMDGP has many interesting properties which hold almost surely:

- G is rigid, so $|X|$ is finite; [27]
- in particular, $|X|$ is a power of two; [31]
- the BP algorithm is Fixed-Parameter Tractable (FPT) on the DMDGP [29], and in all the protein instances we tested, the parameter was always fixed at the same constant, yielding polytime behaviour.

By “almost surely” we mean that the set of weighted input graphs for which the above properties may not hold has Lebesgue measure zero in the set of all weighted input graphs (assuming the weights to be real numbers). The BP algorithm relies on the given distances being precise; unfortunately, however, protein data measured through Nuclear Magnetic Resonance (NMR) is subject to experimental errors, modelled as real intervals $[L, U]$ assigned to all edges $\{u, v\}$ whenever $v - u > 2$ in the vertex order. To overcome this difficulty, two research directions have been pursued: (i) the discretization of the uncertainty intervals [22]; (ii) the analytical description, using Clifford algebra, of the locus of vertex v when the edge $\{v, v - 3\}$ is weighted by an interval [20]. The formulation study in this paper moves a first step towards a third direction: the integration of purely continuous techniques within mixed-combinatorial algorithms such as BP.

This brings us to the *interval* DISTANCE GEOMETRY PROBLEM (*i*DGP), which is a variant of the DGP defined as follows: the edge function is an interval function $[L, U] : E \rightarrow \mathbb{IR}_+$, where L, U are two nonnegative functions from $E \rightarrow \mathbb{R}_+$ such that $L_{uv} \leq U_{uv}$ for each $\{u, v\} \in E$, \mathbb{IR}_+ is the set of nonnegative real intervals, and Eq. (1) is replaced by:

$$\forall \{u, v\} \in E \quad L_{uv} \leq \|x_u - x_v\|_2 \leq U_{uv}. \quad (2)$$

Note that Eq. (2) is often written as:

$$\forall \{u, v\} \in E \quad L_{uv}^2 \leq \|x_u - x_v\|_2^2 \leq U_{uv}^2. \quad (3)$$

As explained later, Eq. (3) minimizes the chances that numerical solvers, which rely on the floating-point representation of real numbers, might stumble upon a negative representation of zero, thereby raising a “not a number” (NaN) error upon calculating the square root.

Most solution techniques for solving *i*DGP instances require a continuous search in Euclidean space, even if the given graph is rigid. The most direct approach is to formulate the *i*DGP as a Mathematical Program (MP), which can then be solved by a MP solver. The aim of this paper is to determine the best solver+formulation combination for the *i*DGP. To this end, we need to know: (a) how to evaluate the quality of the solutions computed by the solvers; (b) which formulations to employ; (c) which solvers to employ. We therefore introduce new and existing error measures, formulations and solvers, before proceeding to evaluate them all computationally.

Several innovations are introduced in this paper: (i) the root square deviation modulo isomers; (ii) some new MP formulations for the *i*DGP; (iii) the concept of “pointwise formulation” to be used in

alternating-type algorithms; (iv) an adaptation of the Multiplicative Weights Update (MWU) algorithm to the *i*DGP. We conclude that the MWU algorithm with its pointwise formulation is the best combination, and that the new “square factoring” MP formulation, used within either a pure MultiStart (MS) or a Variable Neighbourhood Search (VNS) heuristic, is second best.

The rest of the paper is organized as follows. In Sect. 2, we define error measures to meaningfully compare protein backbones found algorithmically with those stored in Protein Data Bank (PDB) files [9], and introduce a new root mean square deviation type measure modulo certain partial reflection isomers. In Sect. 3, we list several formulations, relaxations and variants for the *i*DGP, some of which are new. In Sect. 4, we propose a new algorithm for solving the *i*DGP: namely, an adaptation of the Multiplicative Weights Update method [4]. In Sect. 5, we discuss comparative computational results, which show that, on average, our newly proposed algorithm provides the best quality solutions.

2 Error measures for realizations of protein graphs

Since we aim at ascertaining which formulation(s) can provide the best and/or fastest bound, we need a method to benchmark quality and speed with respect to any solution algorithm. We benchmark speed by simply measuring CPU time.

Benchmarking solution quality is more complicated. Since YES instances of the DGP and *i*DGP can in general yield realizations with irrational components, replaced in computer calculations by floating point numbers, we do not necessarily obtain clear-cut YES/NO answers. One might speak instead of “feasible” or “infeasible” realizations for a certain instance, and even of “good” or “bad” ones relative to some given target realization. Accordingly, we must perform a post-processing analysis of the numerical errors associated to computational experiments in order to establish whether the given instance is actually YES or NO with some error tolerance, and whether the obtained solution is practically meaningful.

2.1 The edge error

Given a realization $x^* : V \rightarrow \mathbb{R}^K$, we can measure the error of x^* with respect to a given *i*DGP instance by assigning an ℓ_2 -norm error to each edge $\{u, v\}$ of the graph $G = (V, E)$, given by [30]:

$$\alpha_{uv}(x^*) = \max(0, L_{uv} - \|x_u^* - x_v^*\|_2) + \max(0, \|x_u^* - x_v^*\|_2 - U_{uv}). \quad (4)$$

We remark that the corresponding error for non-interval DGP instances is:

$$\beta_{uv}(x^*) = \left| \|x_u^* - x_v^*\|_2 - U_{uv} \right|.$$

Accordingly, we define the edge error as follows:

$$\eta_{uv}(x^*) = \begin{cases} \alpha_{uv}(x^*) & \text{if the instance is } i\text{DGP} \\ \beta_{uv}(x^*) & \text{if the instance is DGP.} \end{cases}$$

We can now define the *average error* associated to the instance graph G and a realization x^* as:

$$\Phi(x^*, G) = \frac{1}{|E|} \sum_{\{u,v\} \in E} \eta_{uv}(x^*), \quad (5)$$

and the *maximum error* as:

$$\Psi(x^*, G) = \max_{\{u,v\} \in E} \eta_{uv}(x^*). \quad (6)$$

2.2 The root mean square deviation

The edge errors go a long way in determining when a realization x^* is not valid. In many applications, however, we know *a priori* that a problem instance should be feasible. Take e.g. the reconstruction of protein conformations from inter-atomic distances: the protein certainly exists (this is also the case when localizing sensors in wireless networks: the network is being measured, so it exists). Furthermore, we might have an idea of an approximate realization \bar{x} : in the case of proteins, we might know the “uncoiled” shape, for example, or a different isomer of the protein. In this setting, we want to evaluate the error with respect to the given realization \bar{x} .

An obvious way to adapt the edge error to this situation is to compute the average, over edges in E , of an absolute ℓ_2 -norm distance difference:

$$\Delta(x^*, \bar{x}) = \frac{1}{|E|} \sum_{\{u,v\} \in E} | \|x_u^* - x_v^*\|_2 - \|\bar{x}_u - \bar{x}_v\|_2 |. \quad (7)$$

Unfortunately this approach is wrong, since different congruent realizations yield different error values, making the comparison impossible.

To this end, the *root mean square deviation* (RMSD) is often used instead: i.e., translate both x^* and \bar{x} so that their centroids $\gamma(x^*) = \gamma(\bar{x}) = 0$, where the *centroid* is the vector $\gamma(x) \in \mathbb{R}^K$ defined as:

$$\gamma(x) = \sum_{v \leq K} x_v, \quad (8)$$

and then find the congruence ρ (consisting of a rotation composed with at most one reflection) such that $\|x^* - \rho(\bar{x})\|$ is minimum. Note that the norm $\|\cdot\|$ on \mathbb{R}^{K^n} is induced by the ℓ_2 -norm in \mathbb{R}^K :

$$\|x^* - \bar{x}\| = \sum_{v \in V} \|x_v^* - \bar{x}_v\|_2. \quad (9)$$

The RMSD between x^* and \bar{x} is defined as $\min_{\rho} \|x^* - \rho(\bar{x})\|$.

2.3 Distance error modulo isometries

Although the RMSD is widely used in computational geometry, it still falls short in one of the properties of molecules, namely *isomers*, which are molecules having the same chemical formula but different 3D structure.

If we consider protein backbones only, their graphs $G = (V, E)$ possess a further structural property. They have an order $<$ on V such that:

1. the first K vertices in the order form a clique in G (*clique property*);
2. each vertex $v > K$ is adjacent to $v - 1, \dots, v - K$ (*contiguous trilateration order property*).

Although protein backbones have $K = 3$, we develop the theory for general K . DGP instances having these properties are also collectively known as K DMDGP, which are a subclass of Henneberg type I graphs [19]. Contiguous trilateration orders are also known as cTOP or K DMDGP orders [12]. The edges induced by these properties in a K DMDGP graph are called *discretization edges*, and the edges which are not discretization edges are called *pruning edges*.

Many mathematical aspects of the K DMDGP have been investigated in the past (see [31, 29, 26]). The problem itself is **NP**-hard. The automorphism group of X generally contains a subgroup \mathcal{G}_P consisting of partial reflections g_v , called the *pruning group*, such that the action of g_v over a realization $x \in X$ is:

$$g_v(x) = (x_1, \dots, x_{v-1}, R_x^v(x_v), \dots, R_x^v(x_n)), \quad (10)$$

where R_x^v is the reflection with respect to the affine subspace spanned by x_{v-1}, \dots, x_{v-K} , and where v ranges over a vertex set

$$Z = V \setminus \bigcup_{\substack{\{u,w\} \in E \\ u+K < w}} \{u+K+1, \dots, w\},$$

or, in other words, v must not be “covered” by any pruning edge.

The protein backbone isomers of a valid realization \bar{x} are given by the orbit $\mathcal{G}_P x = \{g_v(x) \mid v \in Z\}$. It turns out that all backbone isomers in $\mathcal{G}_P x$ are valid realizations of the given DGP instance G . So we might obtain a realization x^* which is a valid isomer (and hence has zero edge errors), but has a large RMSD with the given (different) isomer \bar{x} .

A serious issue arises when considering i DGP instances, however: if the RMSD between x^* and \bar{x} is positive, is it due to the “slack” induced by the interval edge weights, or is it due to the fact that x^* and \bar{x} are different isomers of essentially the same backbone (a similar issue was described in [35])? This motivates us to define the following problem:

DISTANCE ERROR MODULO ISOMETRIES (DEMI). Given integers n, K with $n \geq K$, two n -point realizations $x, y \in \mathbb{R}^{Kn}$ such that $\gamma(x) = \gamma(y) = 0$, and a description of a pruning group \mathcal{G}_P , find the rotation ρ and a partial reflection composition $g \in \mathcal{G}_P$ such that $\|x - g\rho(y)\|$ is minimum.

We let $\partial(x, y)$ be the minimum value of $\|x - g\rho(y)\|$ which solves the DEMI. We note that ∂ is *not* a semimetric (hence not even a metric), since $\partial(x, y)$ can be zero even though $x \neq y$ (just take y as a partial reflection of x).

2.3.1 Complexity of DEMI

The computational complexity class of DEMI depends on the description of the pruning group. If it is given explicitly, by listing all the partial reflection compositions in \mathcal{G}_P , then the trivial Algorithm 1 solves the problem in polynomial time for fixed K . For a realization $x \in \mathbb{R}^{Kn}$ and an integer $h \leq n$, let $x[h]$ be the partial realization $(x_i \mid 1 \leq i \leq h)$. Step 1 takes a polynomial amount of time for fixed K (an

Algorithm 1 SolveDEMI(x, y, \mathcal{G}_P)

- 1: Find a congruence ρ minimizing $\|x[K] - y[K]\|$
 - 2: Let $\partial(x, y) = \min\{\|x - g\rho(y)\| \mid g \in \mathcal{G}_P\}$
-

$O(n^{K-2} \log n)$ algorithm was described in [2]), but more efficient methods exist for $K = 3$, see [5, 14]. Step 2 depends linearly on the order of the pruning group, which was shown in [26] to be $2^{|Z|}$. Since Z is usually small in practice (see Sect. 2.3.2) and on average (see Sect. 2.3.3), assuming the input to DEMI to be the explicit list of all partial reflection compositions is not out of place.

We have not been able to prove that DEMI can be solved in polynomial time (for fixed K) if its input is x, n, Z (or any other compact description of the generators of \mathcal{G}_P), nor that DEMI is **NP**-hard under the same conditions. We leave this as an open question.

2.3.2 Empirical observations on the size of Z

In this section we exhibit empirical evidence to the effect that $|Z|$ is rarely large. First, we note that $|Z| \geq 1$: this follows by the definition of $Z = \{v > K \mid \nexists \{u, w\} \in E (u + K < v \leq w)\}$, since $v = K + 1$ is obviously always in Z (this can also be shown by other means [21, Sect. 2.1]).

Figures 1-2 show the mean and standard deviations of $|Z|$ relative to samples of 500 randomly generated K DMDGP instances for each value of $K \in \{2, 3\}$ and various values of the edge sparsity s . The generation procedure is as follows: given $n = |V|$ and K , we initially generate a K DMDGP instance with all the necessary discretization edges in its edge set E (there are $K(K-1)/2 + (n-K)K$ of them), but no pruning edges. Then we loop over all $\{i, j\}$ which are not discretization edges, and with given probability s we insert a pruning edge in E . So s is in fact the density of the pruning edges.

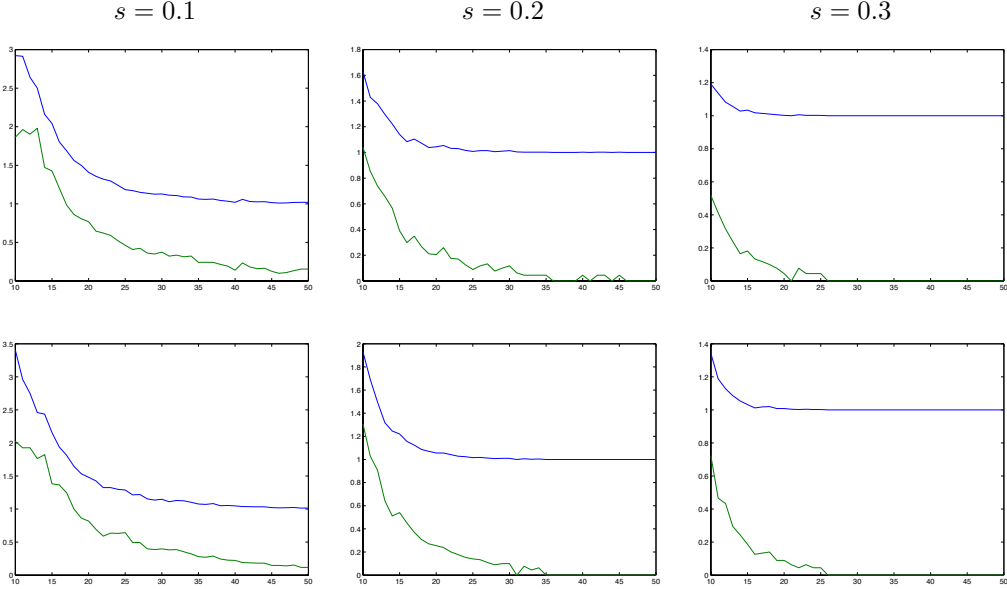


Figure 1: Mean (top curve) and standard deviation (bottom curve) of the pruning group size in function of n for fixed values of $K = 2$ (top), $K = 3$ (bottom), and the edge sparsity s (values in $\{0.1, 0.2, 0.3\}$ in left, middle and right columns).

The exact dependency of $|Z|$ on the number of pruning edges is given in [29], and it is used to show that the BP algorithm is FPT. It should be clear by definition that the denser the graph, the smaller Z must be. Figures 1-2 show (empirically) that $|Z|$ tends to 1 very fast and very reliably as n and s increase, with n, s as small as, respectively, 20 and 0.3. Large graphs with $|Z| > 1$ are very rare.

It is interesting to note that the standard deviation of $|Z|$ in function of the sparsity s has a maximum in $[0, 0.05]$. This phenomenon is analyzed below in more detail.

2.3.3 Expectation and variance of $|Z|$

As explained in Sect. 2.3, K DMDGP instances consist of a backbone subgraph (a minimal graph satisfying the clique and contiguous trilateration order properties) and some pruning edges. Accordingly, random K DMDGP graphs $G = (V, E)$ are generated as follows:

- a backbone which only depends on K, n and determines the order on V ;
- for each pair $\{u, w\}$ which is not a discretization edge, we independently add $\{u, w\}$ as a pruning edge in E with probability $s \in [0, 1]$.

Now consider the subset $Z \subseteq V$, defined as in Sect. 2.3 as

$$Z = \{v > K \mid \nexists \{u, w\} \in E (u + K < v \leq w)\}.$$

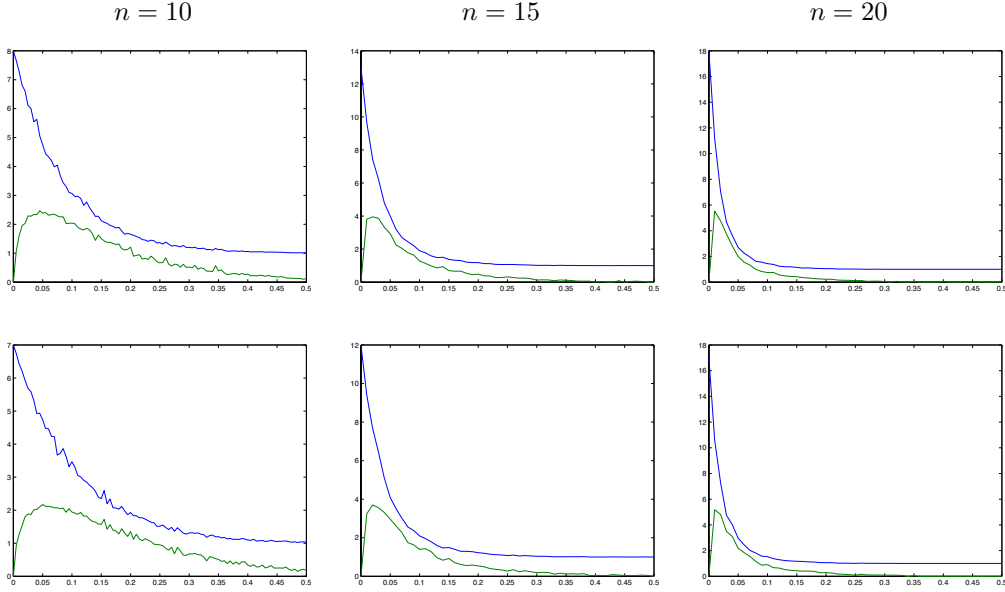


Figure 2: Mean (top curve) and standard deviation (bottom curve) of the pruning group size in function of the edge sparsity s for fixed values of $K = 2$ (top), $K = 3$ (bottom), and n (values in $\{10, 15, 20\}$ in left, middle and right columns).

We consider $|Z|$ as a random variable depending on the edge probability s (also known as the sparsity of the K DMDGP graph G), and compute its expected value. In the following, $P(\cdot)$ is the probability of an event, $E(\cdot)$ is the expectation of a random variable and $\text{Var}(\cdot)$ is its variance.

2.1 Proposition

$$E(|Z|) \leq 1 + (n - K - 1)(1 - s)^{n-K-1}.$$

Proof. For all $v \in \{K + 1, \dots, n\}$ define $\mathcal{X}_v = 0$ if $v \notin Z$ and 1 if $v \in Z$. Then $|Z| = \sum_{v=K+1}^n \mathcal{X}_v$, which implies:

$$E(|Z|) = \sum_{v=K+1}^n E(\mathcal{X}_v) = \sum_{v=K+1}^n P(v \in Z).$$

Now, for any $v \in \{K + 1, \dots, n\}$ there are $v - K - 1$ choices of u with $u + K < v$, and there are $n - v + 1$ choices of w with $v \leq w$. Therefore, there are $(v - K - 1)(n - v + 1)$ possible choices of the pruning edge $\{u, w\}$ such that $u + K < v \leq w$. Moreover, $v \in Z$ if all these pairs are not added to the graph. Thus, $P(v \in Z) = (1 - s)^{(v-K-1)(n-v+1)}$, and hence:

$$E(|Z|) = \sum_{v=K+1}^n (1 - s)^{(v-K-1)(n-v+1)}.$$

Finally, we remark that (a) the first term of the sum is 1, and (b) $(1 - s) < 1$, so we can replace all the terms of the sum by the second largest one, and obtain:

$$E(|Z|) \leq 1 + (n - K - 1)(1 - s)^{n-K-1}, \quad (11)$$

as claimed. \square

The RHS of Eq. (11) converges to 1 as $s \rightarrow 1$ with n, K fixed, and as $n \rightarrow \infty$ with s, K fixed, which is

consistent with the empirical results of Sect. 2.3.2. We are therefore justified in making the qualitative statements that, for random DMGDP, $|Z| \approx 1$.

We now discuss the variance. We have:

$$\begin{aligned}
\text{Var}(|Z|) &= \mathbb{E}(|Z|^2) - (\mathbb{E}(|Z|))^2 \\
&= \mathbb{E}\left(\left(\sum_{v=K+1}^n \mathcal{X}_v\right)^2\right) - \left[\sum_{v=K+1}^n \mathbb{E}(\mathcal{X}_v)\right]^2 \\
&= \sum_{v=K+1}^n \mathbb{E}(\mathcal{X}_v^2) + 2 \sum_{K+1 \leq v_1 < v_2 \leq n} \mathbb{E}(\mathcal{X}_{v_1} \mathcal{X}_{v_2}) \\
&\quad - \sum_{v=K+1}^n [\mathbb{E}(\mathcal{X}_v)]^2 - 2 \sum_{K+1 \leq v_1 < v_2 \leq n} \mathbb{E}(\mathcal{X}_{v_1}) \mathbb{E}(\mathcal{X}_{v_2}) \\
&= \sum_{v=K+1}^n \mathbb{E}(\mathcal{X}_v) + 2 \sum_{K+1 < v \leq n} \mathbb{E}(\mathcal{X}_v) + 2 \sum_{K+1 < v_1 < v_2 \leq n} \mathbb{E}(\mathcal{X}_{v_1} \mathcal{X}_{v_2}) \\
&\quad - \sum_{v=K+1}^n [\mathbb{E}(\mathcal{X}_v)]^2 - 2 \sum_{K+1 < v \leq n} \mathbb{E}(\mathcal{X}_v) - 2 \sum_{K+1 < v_1 < v_2 \leq n} \mathbb{E}(\mathcal{X}_{v_1}) \mathbb{E}(\mathcal{X}_{v_2}) \\
&\quad \text{(this follows from the fact that } \mathbb{E}(\mathcal{X}_{K+1}) = 1 \text{ and } \mathbb{E}(\mathcal{X}_{K+1} \mathcal{X}_v) = \mathbb{E}(\mathcal{X}_v) \text{ for all } v) \\
&= \sum_{v=K+2}^n \mathbb{E}(\mathcal{X}_v) + 2 \sum_{K+2 \leq v_1 < v_2 \leq n} \mathbb{E}(\mathcal{X}_{v_1} \mathcal{X}_{v_2}) \\
&\quad - \sum_{v=K+2}^n [\mathbb{E}(\mathcal{X}_v)]^2 - 2 \sum_{K+2 \leq v_1 < v_2 \leq n} \mathbb{E}(\mathcal{X}_{v_1}) \mathbb{E}(\mathcal{X}_{v_2}).
\end{aligned}$$

By definition of Z , two vertices v_1 and v_2 are in Z if all pairs $\{u, w\}$ such that either $u + K < v_1 \leq w$ or $u + K < v_2 \leq w$ are not edges of G . Assume $v_1 < v_2$, then there are:

$$\begin{aligned}
&(v_1 - K - 1)(n - v_1 + 1) + (v_2 - K - 1)(n - v_2 + 1) - (v_1 - K - 1)(n - v_2 + 1) \\
&= (v_1 - K - 1)(n - v_1 + 1) + (v_2 - v_1)(n - v_2 + 1)
\end{aligned}$$

such edges (by counting all pairs of each type and subtracting the number of doubly counted ones). So, the probability that $v_1, v_2 \in Z$ is

$$(1 - s)^{(v_1 - K - 1)(n - v_1 + 1) + (v_2 - v_1)(n - v_2 + 1)}.$$

This implies

$$\begin{aligned}
\text{Var}(|Z|) &= \sum_{v=K+2}^n (1 - s)^{(v - K - 1)(n - v + 1)} - \sum_{v=K+2}^n (1 - s)^{2(v - K - 1)(n - v + 1)} + \\
&\quad + 2 \sum_{K+2 \leq v_1 < v_2 \leq n} (1 - s)^{(v_1 - K - 1)(n - v_1 + 1) + (v_2 - v_1)(n - v_2 + 1)} - \\
&\quad - 2 \sum_{K+2 \leq v_1 < v_2 \leq n} (1 - s)^{(v_1 - K - 1)(n - v_1 + 1) + (v_2 - K - 1)(n - v_2 + 1)}.
\end{aligned}$$

To simplify the analysis of $\text{Var}(|Z|)$, we provide an upper bound.

2.2 Lemma

For all $s \in (0, 1)$ and $k \geq 1$, we have $\sum_{i=1}^{k-1} (1 - s)^{i(k-i)} < \frac{2(1-s)^{k-1}}{s}$.

Proof. For each $1 \leq i < \lfloor \frac{k}{2} \rfloor$ we have the estimate

$$(i + 1)(k - i - 1) = ik + k - i^2 - 2i - 1 = i(k - i) + k - 2i - 1 \geq i(k - i) + 1. \quad (12)$$

Therefore,

$$\begin{aligned}
\sum_{i=1}^{k-1} (1-s)^{i(k-i)} &\leq 2 \sum_{i=1}^{\lfloor \frac{k}{2} \rfloor} (1-s)^{i(k-i)} \\
&\leq 2((1-s)^{k-1} + (1-s)^k + (1-s)^{k+1} + \dots + (1-s)^{k-2+\lfloor \frac{k}{2} \rfloor}) \\
&< 2(1-s)^{k-1} \sum_{i=0}^{\infty} (1-s)^i \\
&= \frac{2(1-s)^{k-1}}{s}.
\end{aligned}$$

The second inequality follows because of estimate (12). \square

\square

We can now improve the estimate for the variance (where n, K only appear in the exponent):

$$\begin{aligned}
0 &< \text{Var}(|Z|) \\
&< \sum_{v_2=K+2}^n (1-s)^{(v_2-K-1)(n-v_2+1)} + 2 \sum_{K+2 \leq v_1 < v_2 \leq n} (1-s)^{(v_1-K-1)(n-v_1+1)+(v_2-v_1)(n-v_2+1)} \\
&< \sum_{i=1}^{n-K-1} (1-s)^{i(n-K-i)} + 2 \sum_{K+2 \leq v_1 \leq n} \left((1-s)^{(v_1-K-1)(n-v_1+1)} \sum_{i=1}^{n-v_1} (1-s)^{i(n-v_1-i+1)} \right) \\
&< \frac{2}{s} (1-s)^{(n-K-1)} + \frac{4}{s} \sum_{K+2 \leq v_1 \leq n} [(1-s)^{(v_1-K-1)(n-v_1+1)} (1-s)^{n-v_1}] \quad (\text{Lemma 2.2}) \\
&= \frac{2}{s} (1-s)^{(n-K-1)} + \frac{4}{s(1-s)} \sum_{i=2}^{n-K} (1-s)^{i(n-K-i+1)} \\
&< \frac{2}{s} (1-s)^{(n-K-1)} + \frac{4}{s(1-s)} \left(\frac{2}{s} - 1 \right) (1-s)^{n-K} \quad (\text{Lemma 2.2}) \\
&= \left(\frac{8}{s^2} - \frac{2}{s} \right) (1-s)^{n-K-1}.
\end{aligned}$$

For example, with $s = 0.2$, $n = 35$, $K = 2$, the estimate yields $\left(\frac{8}{s^2} - \frac{2}{s} \right) (1-s)^{n-K-1} = 0.15$. With $s = 0.3$, $n = 25$, $K = 2$, we get $\left(\frac{8}{s^2} - \frac{2}{s} \right) (1-s)^{n-K-1} = 0.03$.

Fig. 2 shows that the standard deviation (and hence the variance) of $|Z|$ has a maximum when s is close to zero. Fixing n and K , consider $\text{Var}(|Z|)$ as a function $f(t)$ of $1-s$, let $\tau(k) = kt^k$, and rewrite $\text{Var}(|Z|)$ as:

$$\begin{aligned}
\text{Var}(|Z|) = f(t) &= \sum_{v=K+2}^n t^{(v-K-1)(n-v+1)} - \sum_{v=K+2}^n t^{2(v-K-1)(n-v+1)} + \\
&+ 2 \sum_{K+2 \leq v_1 < v_2 \leq n} t^{(v_1-K-1)(n-v_1+1)+(v_2-v_1)(n-v_2+1)} - \\
&- 2 \sum_{K+2 \leq v_1 < v_2 \leq n} t^{(v_1-K-1)(n-v_1+1)+(v_2-K-1)(n-v_2+1)}.
\end{aligned}$$

Taking the derivative of $f(t)$, we have:

$$f'(t) = t^{-1} \left(\sum_{v=K+2}^n \tau((v-K-1)(n-v+1)) - \sum_{v=K+2}^n \tau(2(v-K-1)(n-v+1)) + \right. \\ \left. + 2 \sum_{K+2 \leq v_1 < v_2 \leq n} \tau((v_1-K-1)(n-v_1+1) + (v_2-v_1)(n-v_2+1)) - \right. \\ \left. - 2 \sum_{K+2 \leq v_1 < v_2 \leq n} \tau((v_1-K-1)(n-v_1+1) + (v_2-K-1)(n-v_2+1)) \right).$$

Consider the derivative of τ with respect to k , $\tau'(k) = (kt^k)' = t^k(1 + k \ln(t))$, and take for example $k \geq 20$ and $t \leq 0.95$. We have $(1 + k \ln(t)) \leq 1 + 20 \ln(0.95) = -0.026 < 0$. Therefore, when $t < 0.95$, $\tau(k)$ is a decreasing function on the set $\{k \mid k \geq 20\}$. It means that, whenever $n - K - 1 \geq 20$, $\tau((v - K - 1)(n - v + 1)) \geq \tau(2(v - K - 1)(n - v + 1))$ for each $v \in \{K + 2, \dots, n\}$, and $\tau((v_1 - K - 1)(n - v_1 + 1) + (v_2 - v_1)(n - v_2 + 1)) \geq \tau((v_1 - K - 1)(n - v_1 + 1) + (v_2 - K - 1)(n - v_2 + 1))$ for each $v_1 < v_2 \in \{K + 2, \dots, n\}$, since all values under τ are at least 20. We therefore have that $f'(t) \geq 0$ for all $t < 0.95$, i.e., whenever $s \in [0.05, 1]$, $\text{Var}(|Z|)$ decreases as s increases. In other words, the maximum of $\text{Var}(|Z|)$ can only be attained on $[0, 0.05]$.

We can generalize this example to the following result.

2.3 Lemma

For fixed n, K , the maximum of $\text{Var}(|Z|)$ can only be attained at $s \in [0, \frac{1}{n-K-1}]$.

Proof. We have

$$\tau'(k) < 0 \Leftrightarrow 1 + k \ln(t) < 0 \Leftrightarrow \ln\left(\frac{1}{t}\right) > \frac{1}{k} \Leftrightarrow \frac{1}{t} > e^{1/k} \Leftrightarrow t < e^{-1/k} \Leftrightarrow s > 1 - e^{-1/k}.$$

Since

$$e^{-1/k} = 1 - \frac{1}{k} + \frac{1}{2!k^2} - \frac{1}{3!k^3} + \dots > 1 - \frac{1}{k},$$

we have $1 - e^{-1/k} < \frac{1}{k}$. Therefore, if $s > \frac{1}{k}$ we have $\tau'(k) < 0$. So, when $s > \frac{1}{n-K-1}$, we have $\tau'(k) < 0$ for all $k \geq n - K - 1$. Now the same argument as in the example above shows that $\text{Var}(|Z|)$ decreases on the set $[\frac{1}{n-K-1}, \infty)$. \square

2.3.4 Computing DEMI measures in practice

We believe we made a convincing argument that we can safely use Alg. 1 to solve DEMI instances. There is, however, a glitch: none of the PDB instances we consider actually comes with a pre-defined cTOP order. For some of them, the protein backbone is a cTOP order. For others this is not the case. The state of the art in automatically finding cTOP orders in graphs is severely limited [12], and certainly does not scale to hundreds of vertices easily. Thus the *DEMI measure* $\partial(x, y)$ of a realization x with respect to a given realization y will not be computed for all instances we test in Sect. 5, but only for some (see Table 10).

3 New and existing *i*DGP formulations

All formulations we consider are box-constrained to bounds $x \in [M^L, M^U]^{Kn}$, which have to be large enough to accommodate a worst-case realization with the given distances. One could take for example

$M^L = -\frac{1}{2} \sum_{\{u,v\} \in E} U_{uv}$ and $M^U = -M^L$, and then tighten these bounds using some pre-processing techniques [7]. We do not write these bounds explicitly in the formulations below. Notationwise, $\mathbf{M} = [M^L, M^U]^m$ and $\mathbf{M}^+ = \mathbf{M} \cap [0, +\infty]$.

Most formulations come with variants. A common variant, which we refer to as the *square root variant*, is the following: replace $\|x_u - x_v\|_2^2$ by $\|x_u - x_v\|_2$ and squared distance bounds by distance bounds. In such variants, because of floating point issues, $\sqrt{\alpha}$ is implemented as $\sqrt{\alpha + \delta}$, where δ is a constant in $O(10^{-10})$.

In all of our formulations, aside from the semidefinite programming (SDP) ones, we fix the centroid at the origin, which means that we find solutions modulo translations. This seems to improve the overall reliability and convergence speed of the heuristic solution algorithms we use. It is interesting that this ceases to be the case if we also impose no rotation by fixing the first K vertices, in which case the algorithms find much worse local optima.

3.1 Validation

With each formulation, we present performances and results on a single PDB instance called `tiny`, which describes a graph $G_{\text{tiny}} = (V, E)$ with $|V| = 37$, $|E| = 335$ and $K = 3$. Fig. 3 shows a heat map of the partial Euclidean Distance Matrix (pEDM) and the correct realization (found in the PDB file) in \mathbb{R}^K using two types of plots.

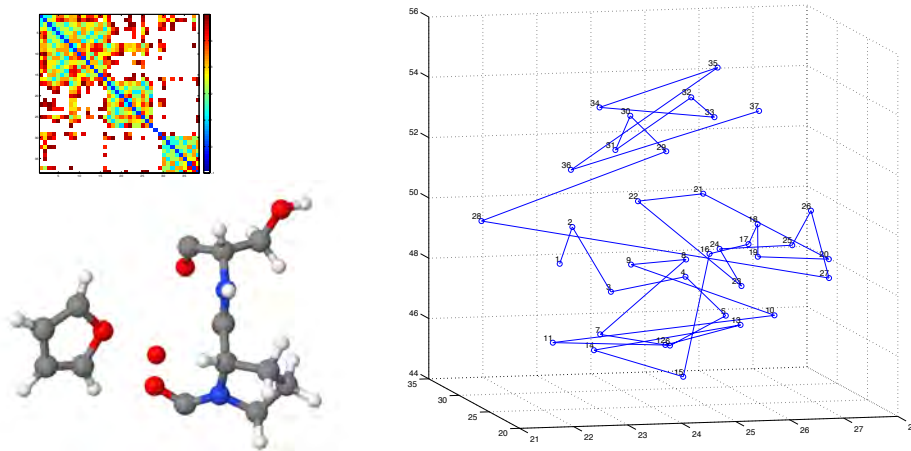


Figure 3: The `tiny` instance: a heat map of the pEDM (upper left) and the correct realization in \mathbb{R}^K shown by the Jmol molecular visualization software (lower left) and in a Euclidean space plot, using the natural vertex order (right). The axes of the two 3D plots are not aligned to minimize overlap. The atom which appears disconnected in the Jmol plot corresponds to vertex 28 in the Euclidean plot, and the nine atoms in a pentagonal arrangement correspond to vertices 29-37.

These validation experiments consist in solving the `tiny` instance using three different Global Optimization (GO) methods. The first method is a deterministic GO solver based on spatial Branch-and-Bound (sBB) [7], which we run for at most 900s. The second method is a stochastic metaheuristic called Variable Neighbourhood Search (VNS), described in [24] with some adaptations from [33]. The third method is a straightforward MultiStart (MS) algorithm, which is possibly the simplest stochastic metaheuristic, and consists of deploying a certain number of local descents from randomly sampled initial points. Both VNS and MS were allowed to run for at most 20s of user CPU time (but terminated when-

ever they found an optimum with average error less than 10^{-6}). The results report the average edge error Φ (see Eq. (5)), the maximum edge error Ψ (see Eq. (6)), the DEMI measure ∂ , and the CPU time in seconds. All statistics referring to stochastic algorithms are averaged over 10 runs.

These validation experiments were conducted on a single core of a two-core Intel i7 CPU running at 2.0GHz with 8GB RAM under the Darwin Kernel v. 13.3.0. Our sBB solver of choice is COUENNE [7] in its default setting. We used AMPL [18] to implement the VNS and MS algorithm, and IPOPT [13, 46] as a local solver. The SDP formulations were modelled using YalmIP [34] running under MATLAB [37] and solved using MOSEK [40].

3.2 Exact formulations

These formulations will yield a valid realization at every global optimum.

3.2.1 Penalty minimization

This formulation minimizes the sum of non-negative penalties s_{uv} deriving from the fact that $\|x_u - x_v\|_2$ is smaller than L_{uv} or larger than U_{uv} :

$$\left. \begin{array}{l} \min_{s \in \mathbf{M}^+, x} \sum_{\{u,v\} \in E} s_{uv} \\ \forall \{u,v\} \in E \quad L_{uv}^2 - \|x_u - x_v\|_2^2 \leq s_{uv} \\ \forall \{u,v\} \in E \quad \|x_u - x_v\|_2^2 - U_{uv}^2 \leq s_{uv} \\ \forall k \leq K \quad \sum_{v \in V} x_{vk} = 0. \end{array} \right\} \quad (13)$$

Variants: (i) replace \sum with \max ; (ii) use different variables s^L, s^U to represent penalties w.r.t. L, U ; (iii) replace the objective by any positive linear form in the penalty variables.

This formulation and its variants have the property that an optimum is global if and only if the objective function value is identically zero. An unconstrained and weighted version of this formulation appeared in [39]. The performance of the penalty minimization formulation and its variants on the `tiny` instance is shown in Table 1.

3.2.2 Square factoring

This formulation has been adapted to the interval case from [15]. It exploits the identity $\|x_u - x_v\|_2^2 = (x_u - x_v)(x_u - x_v)$:

$$\left. \begin{array}{l} \min_{x, \sigma \in \mathbf{M}^K, \tau \in \mathbf{M}^K} \sum_{\{u,v\} \in E} \sum_{k \leq K} (\sigma_{uvk} - \tau_{uvk})^2 \\ \forall \{u,v\} \in E, k \leq K \quad x_{uk} - x_{vk} = \sigma_{uvk} \\ \forall \{u,v\} \in E \quad \sum_{k \leq K} \sigma_{uvk} \tau_{uvk} \geq L_{uv}^2 \\ \forall \{u,v\} \in E \quad \sum_{k \leq K} \sigma_{uvk} \tau_{uvk} \leq U_{uv}^2 \\ \forall k \leq K \quad \sum_{v \in V} x_{vk} = 0. \end{array} \right\} \quad (14)$$

We propose no variants for this formulation. The performance of the square factoring formulation on the `tiny` instance is shown in Table 2.

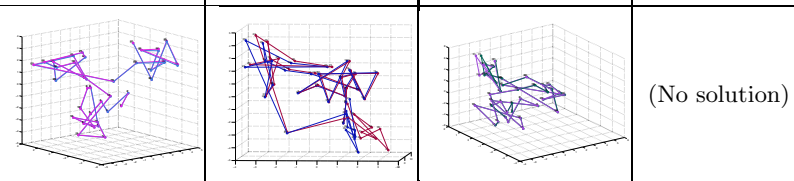
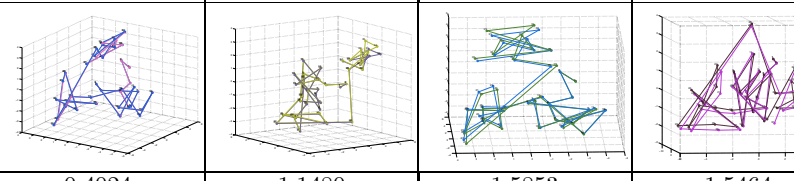
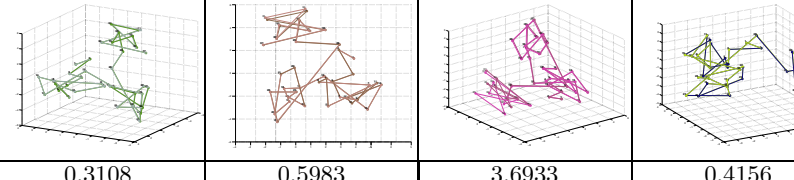
<i>Solver</i>	Original			Var. (i)			Var. (ii)			Var. (iii)		
	Φ	Ψ	<i>CPU</i>	Φ	Ψ	<i>CPU</i>	Φ	Ψ	<i>CPU</i>	Φ	Ψ	<i>CPU</i>
COUENNE	0	0	38.89	0	0	146	0	0	22.05	∞	∞	900
x_{tiny} and x_{DEMI}										(No solution)		
∂	0.3540			2.9834			0.6072			∞		
VNS	0	0	2.27	0.01	0.19	2.76	0.01	0.15	1.09	0	0	3.48
x_{tiny} and x_{DEMI}												
∂	0.4024			1.1480			1.5853			1.5464		
MS	0	0	1.90	0	0	2.28	0	0	1.82	0	0	1.54
x_{tiny} and x_{DEMI}												
∂	0.3108			0.5983			3.6933			0.4156		

Table 1: Performance of *penalty minimization* on *tiny*. For each solver and formulation (variant), we report the edge errors Φ, Ψ , the CPU time, a 3D plot of the solution x_{tiny} given in the PDB file versus the solution x_{DEMI} found by solving the DEMI instance with $x = x_{\text{tiny}}$ and y given by the solution of the solver, and the corresponding DEMI measure $\partial(x, y) = \min_{g, \rho} \|x - g\rho(y)\|$.

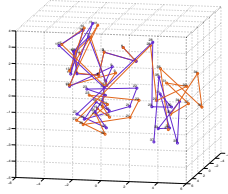
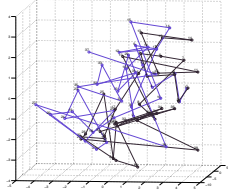
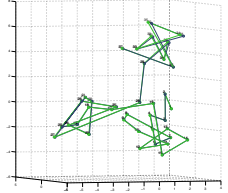
<i>Solver</i>	Φ	Ψ	<i>CPU</i>
COUENNE	0	0	3.11
x_{tiny} and x_{DEMI}			
∂	6.3182		
VNS	0.01	0.02	4.05
x_{tiny} and x_{DEMI}			
∂	13.6406		
MS	0	0	2.31
x_{tiny} and x_{DEMI}			
∂	0.6072		

Table 2: Performance of *square factoring* on `tiny`. For each solver, we report the edge errors Φ, Ψ , the CPU time, a 3D plot of the solution x_{tiny} given in the PDB file versus the solution x_{DEMI} found by solving the DEMI instance with $x = x_{\text{tiny}}$ and y given by the solution of the solver, and the corresponding DEMI measure $\partial(x, y) = \min_{g, \rho} \|x - g\rho(y)\|$.

3.3 Relaxations

These are formulations which relax some feasibility constraints. The obtained solution may or may not be a valid (feasible) solution to the given instance. One should always therefore verify that the solution satisfies (2). On the other hand, if a relaxation is infeasible, then so must be the original *i*DGP instance.

3.3.1 Convexity and concavity

This formulation, adapted to the interval case from [15], exploits the convexity and concavity of the equations in Eq. (3) separately:

$$\left. \begin{array}{l} \max_x \quad \sum_{\{u,v\} \in E} \|x_u - x_v\|_2^2 \\ \forall \{u,v\} \in E \quad \|x_u - x_v\|_2^2 \leq U_{uv}^2 \\ \forall k \leq K \quad \sum_{v \in V} x_{vk} = 0. \end{array} \right\} \quad (15)$$

Variants: replace the objective with a positively weighted version thereof.

Eq. (15) is an exact reformulation (in the sense of [23]) of

$$\min_x \sum_{\{u,v\} \in E} (\|x_u - x_v\|^2 - U_{uv}^2)^2, \quad (16)$$

which is possibly the best known Mathematical Programming (MP) formulation of the (non-interval) DGP so far. That Eq. (16) and Eq. (15) have the same solutions can be intuitively visualized the edges $\{u,v\}$ of the underlying graph G as a set of interconnected cables, each of length U_{uv} : the objective of Eq. (15) “pulls” the adjacent vertices u, v apart as far as possible. As a result, all cables can be straightened if and only if the DGP has a valid solution. A formal proof of this fact is given elsewhere [32].

If the given instance is an *i*DGP one, however, Eq. (15) is a relaxation of the lower bounding constraints: by attempting to maximize the distance between adjacent points, one hopes that $\|x_u - x_v\|_2 \geq L_{uv}$ will hold, but this need not necessarily be the case. The performance of the convexity and concavity formulation and its variants on the `tiny` instance is shown in Table 3.

3.3.2 Semidefinite programming relaxation

This is a natural SDP relaxation, similar to many which already appeared in the literature, where $\|x_u - x_v\|_2^2$ is linearized to $X_{uu} + X_{vv} - 2X_{uv}$:

$$\left. \begin{array}{l} \max_{X \succeq 0} \quad \sum_{\{u,v\} \in E} (X_{uu} + X_{vv} - 2X_{uv}) \\ \forall \{u,v\} \in E \quad X_{uu} + X_{vv} - 2X_{uv} \geq L_{uv}^2 \\ \forall \{u,v\} \in E \quad X_{uu} + X_{vv} - 2X_{uv} \leq U_{uv}^2, \end{array} \right\} \quad (17)$$

where $X \succeq 0$ means that X is required to be positive semidefinite. Several SDP formulations for the DGP have been proposed in the literature over the years, see e.g. [47, 1, 10]. Our formulation, which addresses the *i*DGP, is directly inspired by those in [10], since it employs a linearization of the constraints in Eq. (3). As objective function, we employ a linearization of $\sum_{\{u,v\} \in E} \|x_u - x_v\|^2$, which is unusual. We observed empirically that this yields a good performance on datasets arising from protein conformation.

Variants: replace the objective with $\min \text{Tr}(X)$ as a proxy to rank minimization [11]. The performance of the SDP relaxation and its variant on the `tiny` instance is shown in Table 4.

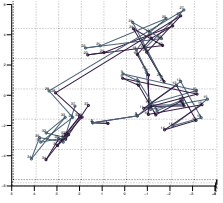
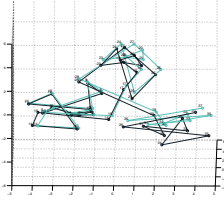
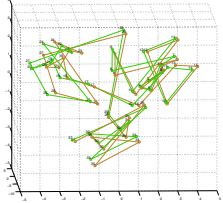
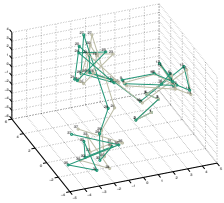
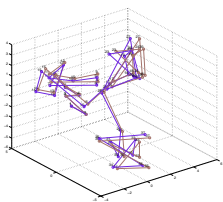
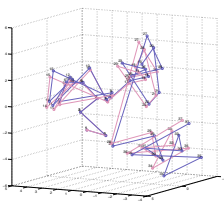
<i>Solver</i>	Original			Var. (i)		
	Φ	Ψ	<i>CPU</i>	Φ	Ψ	<i>CPU</i>
COUENNE	0	0.03	1.78	0	0.03	1.53
x_{tiny} and x_{DEMI}						
∂	2.0211			3.6479		
VNS	0	0.33	8.80	0	0.36	8.94
x_{tiny} and x_{DEMI}						
∂	2.0211			2.3988		
MS	0	0.33	20.19	0	0.35	20.12
x_{tiny} and x_{DEMI}						
∂	2.0211			2.8338		

Table 3: Performance of *convexity and concavity* on *tiny*. For each solver and formulation (variant), we report the edge errors Φ, Ψ , the CPU time, a 3D plot of the solution x_{tiny} given in the PDB file versus the solution x_{DEMI} found by solving the DEMI instance with $x = x_{\text{tiny}}$ and y given by the solution of the solver, and the corresponding DEMI measure $\partial(x, y) = \min_{g, \rho} \|x - g\rho(y)\|$.

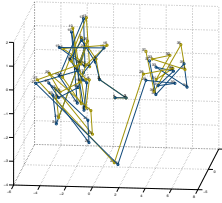
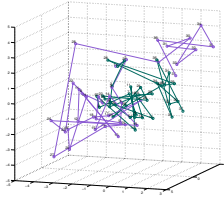
<i>Solver</i>	Original			Var. (i)		
	Φ	Ψ	<i>CPU</i>	Φ	Ψ	<i>CPU</i>
MOSEK	0.0153	0.3140	1.37	0.4704	2.4810	1.35
x_{tiny} and x_{DEMI}						
∂	2.3972			15.6098		

Table 4: Performance of *semidefinite programming* on *tiny*. For each formulation (variant), we report the edge errors Φ, Ψ , the CPU time, a 3D plot of the solution x_{tiny} given in the PDB file versus the solution x_{DEMI} found by solving the DEMI instance with $x = x_{\text{tiny}}$ and y given by the solution of the solver, and the corresponding DEMI measure $\partial(x, y) = \min_{g, \rho} \|x - g\rho(y)\|$.

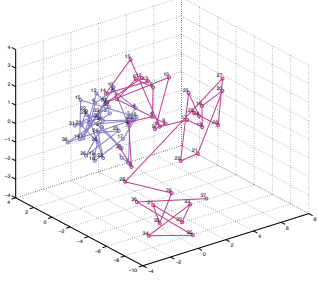
Solver	Original		
	Φ	Ψ	CPU
MOSEK	0.3864	2.6086	1.65
x_{tiny} and x_{DEMI}			
∂	27.3586		

Table 5: Performance of *Yajima's SDP* on *tiny*. We report the edge errors Φ, Ψ , the CPU time, a 3D plot of the solution x_{tiny} given in the PDB file versus the solution x_{DEMI} found by solving the DEMI instance with $x = x_{\text{tiny}}$ and y given by the solution of the solver, and the corresponding DEMI error $\partial(x, y) = \min_{g, \rho} \|x - g\rho(y)\|$.

3.3.3 Yajima's SDP relaxation

This formulation was proposed in [47]. The term $2 \sum_{\{u,v\} \in E} X_{uv}$ added to the objective function is equal to $\text{Tr}(\mathbf{1}X)$ (where $\mathbf{1}$ is the all-one matrix) and has a regularization purpose, ensuring that $\text{Tr}(\mathbf{1}X) = 0$ and hence that $\text{rk}(X) \leq n - 1$.

$$\left. \begin{array}{l} \min_{s \in \mathbf{M}^+, X \succeq 0} \quad \sum_{\{u,v\} \in E} (s_{uv} - (X_{uu} + X_{vv} - 2X_{uv}) + L_{uv}^2) \quad + \quad 2 \sum_{\{u,v\} \in E} X_{uv} \\ \forall \{u, v\} \in E \quad (X_{uu} + X_{vv} - 2X_{uv}) - L_{uv}^2 \leq s_{uv} \\ \forall \{u, v\} \in E \quad 2(X_{uu} + X_{vv} - 2X_{uv}) - L_{uv}^2 - U_{uv}^2 \leq s_{uv} \end{array} \right\} \quad (18)$$

We propose no variants for this formulation. The performance of Yajima's SDP relaxation on the *tiny* instance is shown in Table 5.

3.4 A pointwise reformulation

Pointwise reformulations are only exact for a specific set of values assigned to certain parameters. Typically, replacing variables or entire terms by parameters makes it possible to obtain formulations for which there exist very efficient solution methods. This reformulation will be used in a stochastic search setting (see Sect. 4 below) where the global search phase occurs over the parameter values.

We replace the term $(x_{uk} - x_{vk})^2 = (x_{uk} - x_{vk})(x_{uk} - x_{vk})$ by a linear term $\theta_{uvk}(x_{uk} - x_{vk})$ whenever it occurs in Eq. (3) and (15) in a nonconvex way:

$$\left. \begin{array}{l} \max_x \quad \sum_{\{u,v\} \in E} \sum_{k \leq K} \theta_{uvk}(x_{uk} - x_{vk}) \\ \forall \{u, v\} \in E \quad \|x_u - x_v\|_2^2 \leq U_{uv}^2 \\ \forall \{u, v\} \in E \quad \sum_{k \leq K} \theta_{uvk}(x_{uk} - x_{vk}) \geq L_{uv}^2 \end{array} \right\} \quad (19)$$

It should be clear that for each solution x^* of Eq. (3), there is a parameter matrix $\theta^* \in \mathbb{R}^{mK}$ such that x^* is a feasible solution of Eq. (19): it suffices to choose $\theta_{uvk}^* = (x_{uk}^* - x_{vk}^*)$ for each $\{u, v\} \in E$ and

$k \leq K$. Note that Eq. (19) is a convex MP, and can therefore be solved efficiently. We let $\text{PtwCvx}(\theta)$ be the solution of Eq. (19) with input parameters θ .

4 A new i DGP algorithm

In this section we discuss an adaptation to the i DGP of the well-known MWU method [4]. As explained in [4], the MWU is in fact a meta-algorithm: it has been rediscovered along the years applied to many different optimization problems. Differently from most meta-heuristics, the MWU is as much a theoretical tool as a practical method, insofar as it provides a “generic” asymptotic performance guarantee which works for all problems where the MWU applies. The performance guarantee proof can be modified according to the specific features of the given problem to yield theoretical results. Among the problems listed in [4], possibly the most interesting for the GO community are the Plotkin-Shmoys-Tardos LP feasibility approximation algorithm [42] and the SDP approximation algorithm in [3].

The MWU is applied to a multi-iteration setting over a given horizon $\{1, \dots, T\}$ where, at each iteration $t \leq T$, m “advisors” express an opinion about a certain decision. The advisors’ opinion yield a gain/loss vector $\psi^t = (\psi_i^t \mid i \leq m)$ in $[-1, 1]^m$. The MWU method associates a discrete distribution $\rho^t = (\rho_i^t \mid i \leq m)$ on the advisors, which is updated using the rule

$$\omega_i^t = \omega_i^{t-1}(1 - \eta\psi_i^{t-1}) \quad (20)$$

for each $t > 1$, where $\rho_i^t = \frac{\omega_i^t}{\sum_{\ell} \omega_{\ell}^t}$ and $\eta \leq \frac{1}{2}$ is a user-defined parameter. This distribution essentially measures the reliability of each advisor. The method then stochastically takes the decision given by advisor i with probability ρ_i^t . The average gain/loss made by MWU is therefore given by the weighted average $\Omega^t = \psi^t \cdot \rho^t$. It is shown in [4] that the following bound holds:

$$\sum_{t \leq T} \Omega^t \leq \sum_{t \leq T} \psi_{\ell}^t + \eta \sum_{t \leq T} |\psi_{\ell}^t| + \frac{\ln m}{\eta}, \quad (21)$$

where ℓ is the index of the *best advisor* on average over all iterations. For fixed m and $T \rightarrow \infty$, Eq. (21) states that the cumulative gain/loss made by the MWU method is bounded by a (piecewise) linear function of the gain/loss made by the best advisor, which is somewhat counterintuitive, given that ℓ is not known in advance.

4.1 The MWU method in the i DGP setting

We now reinterpret the MWU method in the setting of the i DGP, which aims to solve the problem via the pointwise reformulation Eq. (19). Consider a loop over T iterations: the convex pointwise reformulation Eq. (19) is solved at each iteration and efficiently yields a candidate realization \bar{x} . This is then refined using \bar{x} as a starting point to a local Nonlinear Programming (NLP) solver applied to the penalty minimization formulation of Eq. (13), which yields a current iterate x .

We now explain how x is used to stochastically update θ at iteration $t \leq T$ along the lines of the MWU method (see the summary in Fig. 4):

- let $(D_{uv}) = (\|x_u - x_v\| \mid u, v \in V)$ be the distance matrix corresponding to x ;
- for each $\{u, v\} \in E$ and $t \leq T$, let:

$$\psi_{uv}^t = \frac{\alpha_{uv}}{\max_{\{w, z\} \in E} \alpha_{wz}} \quad (22)$$

be the relative error of D with respect to $[L, U]$, where α_{uv} is defined in Eq. (4) — note that ψ^t is a scaled edge error vector with every component in $[0, 1]$;

- for each $\{u, v\} \in E$ and $1 < t \leq T$ let

$$\omega_{uv}^t = \omega_{uv}^{t-1}(1 - \eta\psi_{uv}^{t-1}); \quad (23)$$

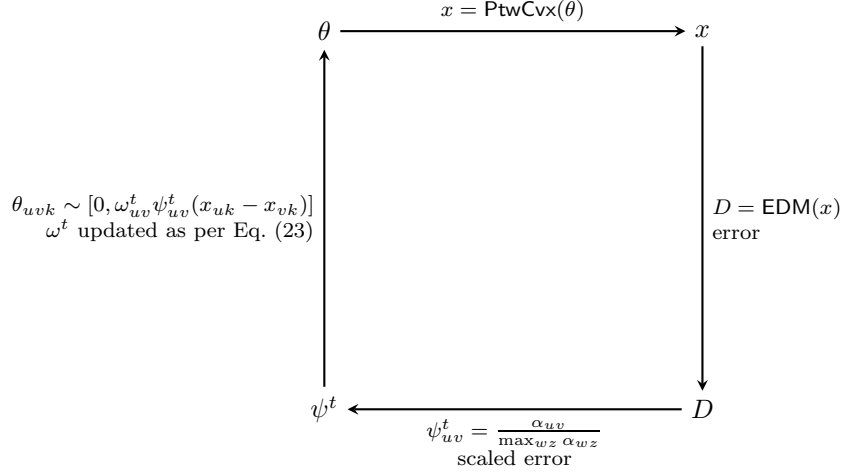


Figure 4: The update of θ from a candidate realization x at each iteration t of the MWU method. The oracle $\text{PtwCvx}(\theta)$ solves the pointwise reformulation Eq. (19) parametrized with θ , and uses the solution as a starting point to a local NLP algorithm solving an exact formulation of the i DGP, say Eq. (13).

- let θ_{uvk} be a random value sampled from the uniform distribution on $[0, \omega_{uv}(x_{uk} - x_{vk})]$.

We remark that the distribution ρ^t is defined in terms of the edge weights ω^t :

$$\rho_{uv}^t = \frac{\omega_{uv}^t}{\sum_{\{w,z\} \in E} \omega_{wz}^t}. \quad (24)$$

The MWU method applied to the i DGP is given as Alg. 2.

Algorithm 2 MULTIPLICATIVEWEIGHTSUPDATE(η, T)

- 1: let $\omega^0 = 1$
 - 2: let x be the output of a local NLP solver applied to Eq. (13)
 - 3: let $x' = x$ be the best solution so far
 - 4: **for** $t \leq T$ **do**
 - 5: derive θ from x as explained above
 - 6: compute a new candidate realization $\bar{x} = \text{PtwCvx}(\theta)$
 - 7: let x be the solution returned by a local NLP solver on Eq. (13) with \bar{x} as starting point
 - 8: if x is an improvement with respect to x' according to the average error Ω^t , let $x = x'$
 - 9: **end for**
-

4.2 The MWU approximation guarantee for the i DGP

One specific feature of the i DGP is that the “advisors” never yield gains but only a cost vector ψ^t having components in $[0, 1]$. This allows us to prove the following result:

4.1 Proposition

After T iterations of the MWU method, the following relationship holds:

$$\min_{t \leq T} \Omega^t \leq \frac{1}{T} \left(\frac{\ln m}{\eta} + (1 + \eta) \min_{\{u,v\} \in E} \sum_{t \leq T} \psi_{uv}^t \right). \quad (25)$$

Proof. By Line 8 in Alg. 2, $\min_{t \leq T} \Omega^t$ is the per-edge error (weighted by the distribution p^t) associated to x' . From Eq. (21), because $\psi_{uv}^t \geq 0$ for all $\{u, v\} \in E, t \leq T$, we get $\psi_{uv}^t = |\psi_{uv}^t|$, whence, by definition of ℓ in Eq. (21):

$$\sum_{t \leq T} \Omega^t \leq (1 + \eta) \min_{\{u,v\} \in E} \sum_{t \leq T} \psi_{uv}^t + \frac{\ln m}{\eta}.$$

Since x' is the realization with lowest error over all $t \leq T$, then $T \min_{t \leq T} \Omega^t \leq \sum_{t \leq T} \Omega^t$, which implies:

$$T \min_{t \leq T} \Omega^t \leq (1 + \eta) \min_{\{u,v\} \in E} \sum_{t \leq T} \psi_{uv}^t + \frac{\ln m}{\eta}.$$

Dividing through by T yields the result. \square

We remark that the RHS of Eq. (21) is the average weighted error of the best realization found by the MWU in T iterations. Prop. 4.1 states that this error is in the order of a linear function of the smallest scaled error (see Eq. (22)) over all edges.

4.3 Pointwise reformulation feasibility

Although the pointwise reformulation is exact for a certain value of θ , it may fail to even be feasible for certain other values of θ . Since this would be an issue for the MWU method, we further relax it to the following (always feasible) form:

$$\left. \begin{array}{l} \max_{x,s} \sum_{\{u,v\} \in E} \left(\sum_{k \leq K} \theta_{uvk} (x_{uk} - x_{vk}) - s_{uv} \right) \\ \forall \{u,v\} \in E \quad \|x_u - x_v\|_2^2 \leq U_{uv}^2 \\ \forall \{u,v\} \in E \quad \sum_{k \leq K} \theta_{uvk} (x_{uk} - x_{vk}) \geq L_{uv}^2 - s_{uv} \\ s \geq 0. \end{array} \right\} \quad (26)$$

5 Computational assessment

The aim of this section is to present results obtained by four solvers (MS, VNS, MWU, and MOSEK) over 19 different formulations, for each of 61 i DGP instances. Since not every solver can be applied to every formulation, and sometimes errors are generated for combinations of solver+formulation with some of the instances, the number of measure vectors is less than $4 \times 19 \times 61$.

5.1 Solver+formulation combinations

More precisely, we apply MS and VNS to Eq. (13) and its 4 variants (the square root variant and 3 explicitly listed ones), Eq. (14) and its square root variant, Eq. (15) and its positively weighted objective

function variant, for a total of 9 formulations. We apply MWU to Eq. (19), and MOSEK to Eq. (17) and its trace variant, and to Eq. (18). We therefore consider 22 different solver+formulation combinations.

Unlike in the validation experiments, we did not consider the sBB solver as most instances are excessively difficult. The rest of the solver set-up is the same. The solvers MS, VNS, MWU, which are all implemented in AMPL, solve NLP subproblems at each iteration using the local NLP solver IPOPT. The SDP formulations were modelled using YalMIP running under MATLAB and solved using MOSEK. Like the validation experiments, all results were obtained on an Intel i7 CPU running at 2.0GHz with 8GB RAM under the Darwin Kernel v. 13.3.0.

5.2 User-configurable parameters

Each of the MP solvers was given at most 20s of user CPU time, excluding the time taken by IPOPT. Each call to IPOPT was also limited to 20s; however, the IPOPT documentation warns that its stopwatch is not checked regularly, but only after certain operations, which on certain instances appear to take place very rarely. This is apparent in Table 9, where many solvers exceed the 20s CPU time limit. MOSEK was given no time limit, since we wanted to find the optimal solution of the SDP.

All tolerances in the AMPL code were set to 1×10^{-6} . IPOPT was used in its default configuration. The VNS maximum neighbourhood radius and the maximum number of local searches deployed in each neighbourhood were both set to 5. The η parameter in MWU was set to 0.5 (its maximum value) after some preliminary testing. MOSEK was used in its default configuration.

5.3 Instances

Instances were obtained from a selection of PDB files by extracting all the atomic coordinates, computing all of the inter-atomic distances, and discarding all those distances exceeding 5\AA . Table 6 reports the instance names, their sizes, and whether they are classified as easy or hard (last column), see Sect. 5.5.

5.4 Weeding out obvious losers

Not every combination of solver and formulation variant is worth considering. Those which find a solution with high average edge error Φ and/or maximum edge error Ψ should be excluded. We proceeded to record Φ , Ψ , and seconds of user CPU time for every combination on every instance, and we computed the average values (over all instances) of Φ , Ψ , and CPU time.

The statistics for the MS, MWU, and VNS solvers are shown in Fig. 5 (more precisely, if μ is an average, we plotted $\log(1 + \mu)$). All variants involving square roots perform really poorly in terms of edge errors. The statistics for the MOSEK solver, limited to instances where $n \leq 200$ because of RAM limitations, are given below.

<i>Formulation</i>	Φ	Ψ	CPU
<code>sdpre1</code>	0.037	0.516	51
<code>sdpre11</code>	0.123	0.678	45
<code>yajima</code>	0.113	0.717	45

The relevant figure in this table is that the SDP relaxation `sdpre1` has much lower average edge error than the other formulations, lower maximum error, and slightly higher CPU time.

These tests show that, on average, the SDP trace variant, Yajima’s relaxation, and all square root variants are not worth considering. The reason why introducing square roots results in performance losses

<i>Instance</i>	$ V $	$ E $	Hard?	<i>Instance</i>	$ V $	$ E $	Hard?
100d	489	5741	H	helix_amber	392	6265	H
1guu-1	150	959	H	labelplot	37	49	E
1guu-4000	150	968	H	lavor11_7-1	11	47	
1guu	150	955	H	lavor11_7-2	11	47	
1PPT	302	3102	H	lavor11_7-b	11	47	
2erl-frag-bp1	39	406		lavor11_7	11	47	
2kxa	177	2711	H	lavor11	11	40	
C0020pdb	107	999	H	lavor30_6-1	30	192	
C0030pk1	198	3247	H	lavor30_6-2	30	202	H*
C0080create.1	60	681	H	lavor30_6-3	30	195	H*
C0080create.2	60	681	H	lavor30_6-4	30	191	H*
C0150alter.1	37	335	H*	lavor30_6-5	30	195	
C0700odd.1	18	39		lavor30_6-6	30	195	
C0700odd.2	18	39		lavor30_6-7	30	195	
C0700odd.3	18	39		lavor30_6-8	30	193	
C0700odd.4	18	39		mdgp4-heuristic	4	6	
C0700odd.5	18	39		mdgp4-optimal	4	6	
C0700odd.6	18	39		names	86	849	H
C0700odd.7	18	39		odd01	18	39	
C0700odd.8	18	39		odd02	36	308	
C0700odd.9	18	39		pept	107	999	H
C0700odd.A	18	39		res_0	108	1410	H
C0700odd.B	18	39		res_1000	108	1506	H
C0700odd.C	36	242		res_2000	108	1404	H
C0700odd.D	36	242		res_2kxa	177	2627	H
C0700odd.E	36	242		res_3000	108	1487	H
C0700odd.F	18	39		res_5000	108	1392	H
C0700.odd.G	36	308		small02	36	242	
C0700.odd.H	36	308		tiny	37	335	H*
cassioli-protein-130731	281	4871	H	water	648	11939	H
GM1_sugar	68	610	H				

Table 6: The test set: 61 PDB instances, their sizes, and their solution hardness.

may be related to the use of the same local subsolver (IPOPT) within all global optimization solver, since it carries out most floating point computations.

A remark about Yajima’s relaxation: although it was introduced specifically for the *i*DGP, it was originally solved using an ad-hoc interior point method. Even though our results show it underperforms on average with respect to MOSEK, this does not negate the (good) results reported in [47].

We call *bad* the solver+formulation combinations we excluded, and *good* the rest. The good combinations are shown below, marked by a “1” in the corresponding entry.

Formulation			Solver			
<i>Description</i>	<i>Notation</i>	<i>Name</i>	<i>MS</i>	<i>MWU</i>	<i>VNS</i>	<i>Mosek</i>
(13)	(13)	Idgp1	1		1	
(13) variant (i)	(13).1	Idgp1var1	1		1	
(13) variant (ii)	(13).2	Idgp1var2	1		1	
(13) variant (iii)	(13).3	Idgp1var3	1		1	
(14)	(14)	Idgp3	1		1	
(15)	(15)	Idgp4	1		1	
(15) variant (i)	(15).1	Idgp4var1	1		1	
(19)	(19)	Imwu		1		
(17)	(17)	sdp1				1

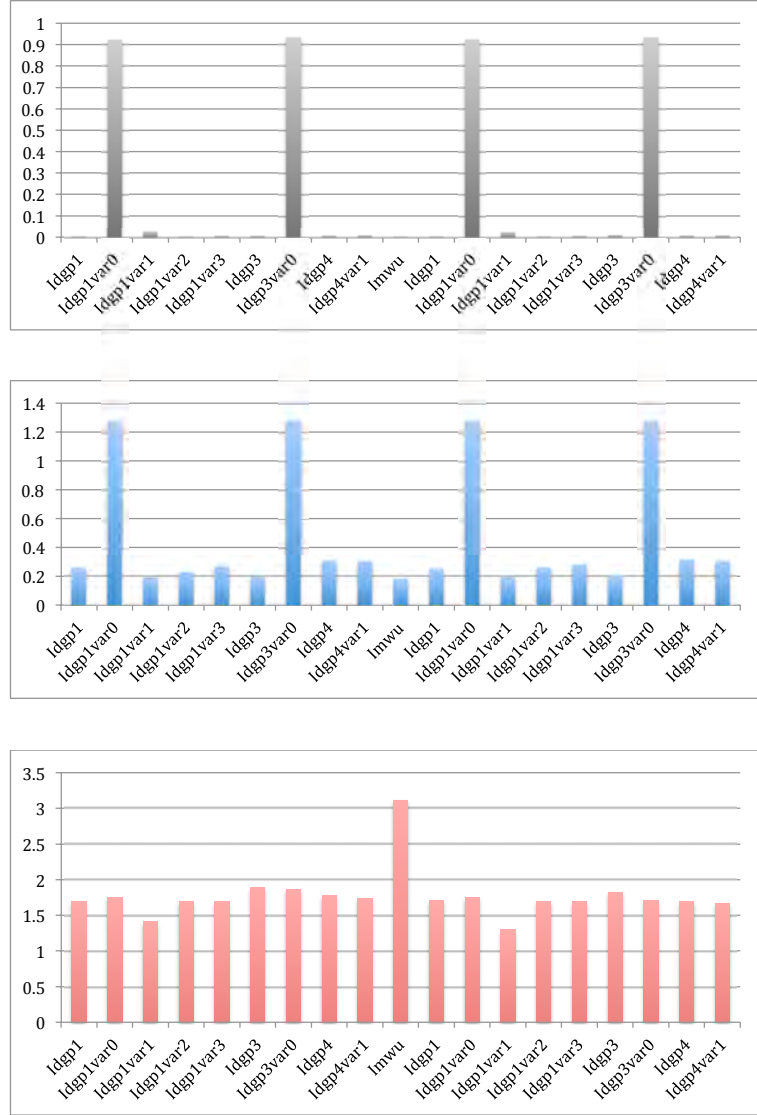


Figure 5: Histogram plots of the statistic $\log(1 + \mu)$ whenever μ is the average of Φ (top), Ψ (middle), and CPU time (bottom) over all instances, for each relevant combination of solver+formulation, with “solver” in MS (left), MWU (middle) and VNS (right).

5.5 Focusing on the hard instances

We also make a qualitative distinction between easy and hard instances. We call an instance *easy* if at least one third of the good combinations find a solution with Φ, Ψ approximately zero within 1s of user CPU time, and *hard* the rest. The classification is reported in the last column of Table 5: hard instances are marked “H”. We marked H* the instances which are “borderline hard”, i.e., there is at least one good combination which finds a solution with Φ, Ψ approximately zero within 1s of user CPU time.

The computational results below will focus on the hard instances for Φ, Ψ ; because of excessive computational requirements, however, we shall relax this constraint for the results on the DEMI measure ∂ .

5.6 Testing heuristics without averages

When benchmarking heuristic algorithms, such as MS, VNS, or MWU, it is customary to present results based on a number of runs (the higher the better) of the same instance. Because of the complexity of this computational comparison, and the absolute time taken to perform it, it was ungainly for us to multiply this effort by a significant factor (say 10 or 100). Does this mean that our results are unreliable? Though it could be argued that the instance-by-instance results are in fact unreliable, as can be gleaned by the difference in ∂ measure for the `tiny` instance in Sect. 3 and those in Table 10, we think the averages (reported in Tables 7-10) are not. Since we never claim in our computational comparisons that one method is best for a certain instance, but only suggest first and second best over all tested instances, we think our computational benchmark is significant.

5.7 Comparative results on edge errors and CPU

In this section we discuss an overarching comparison yielding an overall “winner”. Our most meaningful measure, if Φ and Ψ are nonzero, is the DEMI measure $\partial(\cdot, y)$, where y is a given solution of the ^KDMDGP instance being solved. By Sect. 2.3.4, however, we are not able to compute it for every instance, and hence we focus on Φ, Ψ for our global comparison, and only look at δ on a subset of instances (Sect. 5.8).

Tables 7-9 report the average edge errors Φ , the maximum edge error Ψ , and the CPU time taken by the good combinations when solving hard instances. We remark that the MWU algorithm is best with respect to the edge errors Φ and Ψ , and the worst with respect to CPU time. However, since CPU time is of least consequence in protein conformation computations, CPU time information has a much lower priority than solution quality. We can therefore make the following claim.

The MWU algorithm is the best solver on average.

Given the consequential CPU time difference between MWU and the other solvers, it is worth ranking the solver+formulation combinations by Φ, Ψ , and CPU time (see below).

Rank	Φ		Ψ		CPU	
1	mwu+Imwu	0.029	mwu+Imwu	1.111	vns+Idgp1var1	41.53
2	ms+Idgp1var2	0.031	ms+Idgp1var1	1.237	ms+Idgp1var1	55.11
3	vns+Idgp1	0.032	vns+Idgp1var1	1.259	ms+Idgp4var1	96.85
4	vns+Idgp1var2	0.032	ms+Idgp3	1.265	vns+Idgp4var1	99.05
5	ms+Idgp1	0.033	mosek+sdpre1	1.267	vns+Idgp4	105.70
6	vns+Idgp1var3	0.045	vns+Idgp3	1.281	ms+Idgp1var3	107.41
7	ms+Idgp4	0.046	ms+Idgp1var2	1.560	vns+Idgp1var3	108.15
8	ms+Idgp1var3	0.047	vns+Idgp1	1.752	ms+Idgp1var2	108.63
9	vns+Idgp4	0.047	ms+Idgp1	1.828	ms+Idgp1	109.26
10	ms+Idgp3	0.048	vns+Idgp1var2	1.849	vns+Idgp1var2	109.83
11	vns+Idgp4var1	0.049	ms+Idgp1var3	1.939	ms+Idgp4	111.16
12	ms+Idgp4var1	0.052	vns+Idgp1var3	2.007	vns+Idgp1	113.35
13	vns+Idgp3	0.064	vns+Idgp4var1	2.027	vns+Idgp3	146.98
14	mosek+sdpre1	0.078	ms+Idgp4var1	2.028	ms+Idgp3	170.79
15	vns+Idgp1var1	0.129	ms+Idgp4	2.064	mosek+sdrel	472.82
16	ms+Idgp1var1	0.147	vns+Idgp4	2.140	mwu+Imwu	27669

This ranking shows that `mwu+Imwu` has the only consistent ranking in both Φ and Ψ . It also shows that no other solver+formulation combination has the same desirable property of approximately equal rank w.r.t. both Φ and Ψ . The issue is not only relative: values of Φ higher than 0.1 and of Ψ higher than 1.5 may well imply that the realization is fundamentally wrong, and the only combinations with $\Phi < 0.1$ and $\Psi < 1.5$ are `ms+Idgp3`, `vns+Idgp3`, `mosek+sdpre1`. However, the statistics for the latter

were computed on a subset of instances (all those with $n \leq 200$) due to the high RAM requirements of MOSEK when applied to large instances (see Sect. 5.4). Based on these observations, we claim that:

the formulation `Idgp3` in Eq. (14), when used with MS or VNS, is second best.

We observe that the usual trade-off between quality and efficiency is also at play: solving Eq. (14) takes longest over all formulations solved by both MS and VNS.

5.8 Results on DEMI

Table 10 reports the results on the DEMI measure. Note that the instances in the test set are not the same as for the tests on Φ , Ψ , and CPU (Tables 7-9). As mentioned in Sect. 2.3.4, it is not always possible to determine a cTOP order automatically (or disprove that one exists) in acceptable amounts of CPU time, which is a requirement for computing the DEMI measure. Table 10 includes all instances for which this task could be carried out within 150s of CPU time.

Although it is clear that the SDP relaxation Eq. (17) scores the best performance in terms of the DEMI measure, we mentioned above that the MOSEK solver is unable to scale up to desired sizes. We must therefore resort to the second best, which happens to be the MWU algorithm, consistently with Sect. 5.7. We also observe that VNS attains lower average DEMI measure values more often than MS.

We recall that the DEMI measure values for `tiny` differ from those given in Sect. 3 for the reasons given in Sect. 5.6.

6 Conclusion

Our main aim is to find the best general-purpose continuous search methods for solving *i*DGP instances. To answer this question, we need: (i) a set of benchmarking measures; (ii) a set of *i*DGP formulations; (iii) a set of methods; (iv) extensive computational results. Since a preliminary study [25] showed that two standard metaheuristics and the existing benchmark measures were insufficient, we decided to introduce a new measure and a new method.

Accordingly, this paper presents several notions: (a) a “root mean square deviation modulo partial reflections” (called “DEMI measure”), for benchmarking the performance of *i*DGP algorithms on protein isomers; (b) a zoo of mathematical programming formulations for the *i*DGP; (c) a new method for solving the *i*DGP, based on the well-known Multiplicative Weights Update (MWU) algorithm; (d) a complex computational benchmark for the best formulation-based methods on the hardest instances.

Our study shows that, on average:

- the new MWU-based heuristic yields *i*DGP solutions of highest quality with respect to existing measures;
- the Square Factoring formulation in Eq. (14) is second best;
- as concerns the new DEMI measure, the SDP relaxation in Eq. (17) is best, but only on a limited set of instances, whereas the MWU-based heuristic is second best.

Acknowledgments

The second author (VKK) is supported by a Microsoft Research PhD Fellowship. The third author (CL) is grateful to the Brazilian funding agencies FAPESP and CNPq for financial support. The fourth author (LL) is partly supported by the ANR grant “Bip:Bip” under contract ANR-10-BINF-0003. The fifth author (NM) is grateful to the Brazilian funding agencies FAPERJ and CNPq for financial support.

References

- [1] A. Alfakih, A. Khandani, and H. Wolkowicz. Solving Euclidean distance matrix completion problems via semidefinite programming. *Computational Optimization and Applications*, 12:13–30, 1999.
- [2] H. Alt, K. Mehlhorn, H. Wagnen, and E. Welzl. Congruence, similarity and symmetries of geometric objects. *Discrete Computational Geometry*, 3:237–256, 1988.
- [3] S. Arora, E. Hazan, and S. Kale. Fast algorithms for approximate semidefinite programming using the multiplicative weights update method. In *Foundations of Computer Science*, volume 46 of *FOCS*, pages 339–348. IEEE, 2005.
- [4] S. Arora, E. Hazan, and S. Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8:121–164, 2012.
- [5] M. Atkinson. An optimal algorithm for geometrical congruence. *Journal of Algorithms*, 8:159–172, 1987.
- [6] A. Bahr, J. Leonard, and M. Fallon. Cooperative localization for autonomous underwater vehicles. *International Journal of Robotics Research*, 28(6):714–728, 2009.
- [7] P. Belotti, J. Lee, L. Liberti, F. Margot, and A. Wächter. Branching and bounds tightening techniques for non-convex MINLP. *Optimization Methods and Software*, 24(4):597–634, 2009.
- [8] R. Benedetti and J.-J. Risler. *Real algebraic and semi-algebraic sets*. Hermann, Paris, 1990.
- [9] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I.N. Shindyalov, and P. Bourne. The protein data bank. *Nucleic Acid Research*, 28:235–242, 2000.
- [10] P. Biswas. *Semidefinite programming approaches to distance geometry problems*. PhD thesis, Stanford University, 2007.
- [11] E. Candès, T. Strohmer, and V. Voroninski. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2012.
- [12] A. Cassioli, O. Günlük, C. Lavor, and L. Liberti. Discretization vertex orders for distance geometry. *Discrete Applied Mathematics*, 197:27–41, 2015.
- [13] COIN-OR. *Introduction to IPOPT: A tutorial for downloading, installing, and using IPOPT*, 2006.
- [14] E. Coutsias, C. Seok, and K. Dill. Using quaternions to calculate rmsd. *Journal of Computational Chemistry*, 25(15):1849–1857, 2004.
- [15] C. D’Ambrosio, Vu Khac Ky, C. Lavor, L. Liberti, and N. Maculan. Computational experience on distance geometry problems 2.0. In L. Casado, I. Garcia, and E. Hendrix, editors, *Mathematical and applied Global Optimization*, volume XII of *Global Optimization Workshop*, pages 97–100, Malaga, 2014. University of Malaga.
- [16] Y. Ding, N. Krislock, J. Qian, and H. Wolkowicz. Sensor network localization, Euclidean distance matrix completions, and graph realization. *Optimization and Engineering*, 11:45–66, 2010.

- [17] H. Du, N. Alechina, K. Stock, and M. Jackson. The logic of NEAR and FAR. In T. Tenbrink et al., editor, *COSIT*, volume 8116 of *LNCS*, pages 475–494, Switzerland, 2013. Springer.
- [18] R. Fourer and D. Gay. *The AMPL Book*. Duxbury Press, Pacific Grove, 2002.
- [19] L. Henneberg. *Die Graphische Statik der starren Systeme*. Teubner, Leipzig, 1911.
- [20] C. Lavor, R. Alves, W. Figuereido, A. Petraglia, and N. Maculan. Clifford algebra and the discretizable molecular distance geometry problem. *Advances in Applied Clifford Algebras*, to appear.
- [21] C. Lavor, L. Liberti, N. Maculan, and A. Mucherino. The discretizable molecular distance geometry problem. *Computational Optimization and Applications*, 52:115–146, 2012.
- [22] C. Lavor, L. Liberti, and A. Mucherino. The *interval* Branch-and-Prune algorithm for the discretizable molecular distance geometry problem with inexact distances. *Journal of Global Optimization*, 56:855–871, 2013.
- [23] L. Liberti. Reformulations in mathematical programming: Definitions and systematics. *RAIRO-RO*, 43(1):55–86, 2009.
- [24] L. Liberti and M. Dražić. Variable neighbourhood search for the global optimization of constrained NLPs. In *Proceedings of GO Workshop, Almeria, Spain, 2005*.
- [25] L. Liberti and C. Lavor. Solving large-scale distance geometry problems exactly versus approximately. In *Optimization Society*, Proceedings of the Annual Conference, Houston, 2014. INFORMS.
- [26] L. Liberti, C. Lavor, J. Alencar, and G. Abud. Counting the number of solutions of k DMDGP instances. In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information*, volume 8085 of *LNCS*, pages 224–230, New York, 2013. Springer.
- [27] L. Liberti, C. Lavor, and N. Maculan. A branch-and-prune algorithm for the molecular distance geometry problem. *International Transactions in Operational Research*, 15:1–17, 2008.
- [28] L. Liberti, C. Lavor, N. Maculan, and A. Mucherino. Euclidean distance geometry and applications. *SIAM Review*, 56(1):3–69, 2014.
- [29] L. Liberti, C. Lavor, and A. Mucherino. The discretizable molecular distance geometry problem seems easier on proteins. In Mucherino et al. [41].
- [30] L. Liberti, C. Lavor, A. Mucherino, and N. Maculan. Molecular distance geometry methods: from continuous to discrete. *International Transactions in Operational Research*, 18:33–51, 2010.
- [31] L. Liberti, B. Masson, C. Lavor, J. Lee, and A. Mucherino. On the number of realizations of certain Henneberg graphs arising in protein conformation. *Discrete Applied Mathematics*, 165:213–232, 2014.
- [32] L. Liberti and L. Mencarelli. A multiplicative weights update algorithm for MINLP, 2014. Working paper.
- [33] L. Liberti, N. Mladenović, and G. Nannicini. A recipe for finding good solutions to MINLPs. *Mathematical Programming Computation*, 3:349–390, 2011.
- [34] J. Löfberg. YALMIP: A toolbox for modeling and optimization in MATLAB. In *Proceedings of the International Symposium of Computer-Aided Control Systems Design*, volume 1 of *CACSD*, Taipei, 2004. IEEE.
- [35] V. Maiorov and G. Crippen. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *Journal of Molecular Biology*, 235:625–634, 1994.
- [36] T. Malliavin, A. Mucherino, and M. Nilges. Distance geometry in structural biology. In Mucherino et al. [41].

- [37] The MathWorks, Inc., Natick, MA. *MATLAB R2014a*, 2014.
- [38] J. Milnor. *Topology from the differentiable viewpoint*. University Press of Virginia, Charlottesville, 1969.
- [39] J. Moré and Z. Wu. Distance geometry optimization for protein structures. *Journal of Global Optimization*, 15:219–234, 1999.
- [40] Mosek ApS. *The mosek manual, Version 7 (Revision 114)*, 2014.
- [41] A. Mucherino, C. Lavor, L. Liberti, and N. Maculan, editors. *Distance Geometry: Theory, Methods, and Applications*. Springer, New York, 2013.
- [42] S. Plotkin, D. Shmoys, and É. Tardos. Fast approximation algorithm for fractional packing and covering problems. *Mathematics of Operations Research*, 20:257–301, 1995.
- [43] J. Saxe. Embeddability of weighted graphs in k -space is strongly **NP**-hard. *Proceedings of 17th Allerton Conference in Communications, Control and Computing*, pages 480–489, 1979.
- [44] A. Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and Computational Harmonic Analysis*, 30:20–36, 2011.
- [45] T.-S. Tay and W. Whiteley. Generating isostatic frameworks. *Structural Topology*, 11:21–69, 1985.
- [46] A. Wächter and L. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.
- [47] Y. Yajima. Positive semidefinite relaxations for distance geometry problems. *Japan Journal of Industrial and Applied Mathematics*, 19:87–112, 2002.

<i>Instance</i>	<i>msk</i> <i>IdgPv1</i>	<i>ms</i> <i>IdgPv1</i>	<i>msk</i> <i>IdgPv2</i>	<i>ms</i> <i>IdgPv2</i>	<i>msk</i> <i>IdgPv3</i>	<i>ms</i> <i>IdgPv3</i>	<i>msk</i> <i>IdgPv4</i>	<i>ms</i> <i>IdgPv4</i>	<i>msk</i> <i>IdgPv5</i>	<i>ms</i> <i>IdgPv5</i>	<i>msk</i> <i>IdgPv6</i>	<i>ms</i> <i>IdgPv6</i>	<i>msk</i> <i>IdgPv7</i>	<i>ms</i> <i>IdgPv7</i>	<i>msk</i> <i>IdgPv8</i>	<i>ms</i> <i>IdgPv8</i>	<i>msk</i> <i>IdgPv9</i>	<i>ms</i> <i>IdgPv9</i>	<i>msk</i> <i>IdgPv10</i>	<i>ms</i> <i>IdgPv10</i>	<i>msk</i> <i>IdgPv11</i>	<i>ms</i> <i>IdgPv11</i>	<i>msk</i> <i>IdgPv12</i>	<i>ms</i> <i>IdgPv12</i>	<i>msk</i> <i>IdgPv13</i>	<i>ms</i> <i>IdgPv13</i>	<i>msk</i> <i>IdgPv14</i>	<i>ms</i> <i>IdgPv14</i>	<i>msk</i> <i>IdgPv15</i>	<i>ms</i> <i>IdgPv15</i>	<i>msk</i> <i>IdgPv16</i>	<i>ms</i> <i>IdgPv16</i>	<i>msk</i> <i>IdgPv17</i>	<i>ms</i> <i>IdgPv17</i>	<i>msk</i> <i>IdgPv18</i>	<i>ms</i> <i>IdgPv18</i>				
1004																																								
1PP7																																								
fgur-1																																								
fgur-6000																																								
zxxa																																								
CD20pab																																								
CD68pab																																								
CD68create.1																																								
CD68create.2																																								
CD150alter.1																																								
GH_sugar																																								
castoli...																																								
hellxamber																																								
lavor30_F-2																																								
lavor30_F-3																																								
lavor30_F-4																																								
names																																								
pepc																																								
res-J																																								
res-L000																																								
res-L000																																								
res-zxxa																																								
res-2000																																								
res-3000																																								
res-5000																																								
tiay																																								
water																																								
Average																																								

Table 7: Φ statistics: blue indicates best.

Instance	monet		ms		ms		ms		ms		ms		ms		ms		ms		ms	
	supra	el	IdgP1	IdgP2	IdgP3	IdgP4	IdgP5	IdgP6	IdgP7	IdgP8	IdgP9	IdgP10	IdgP11	IdgP12	IdgP13	IdgP14	IdgP15	IdgP16	IdgP17	IdgP18
1004	NA	3.41	2.10	2.87	3.45	3.45	3.45	2.66	3.03	3.85	3.45	3.85	4.01	3.7	3.08	3.45	3.85	3.78	3.78	3.81
1PPT	NA	0.332	0.0167	0.0667	0.965	2.12	2.66	2.66	2.66	3.76	3.76	3.76	3.76	0.0629	0.00	2.70	2.70	2.70	2.70	3.71
lgau-1	3.2	0.304	0.233	0.45	0.83	0.17	0.17	0.143	0.36	0.677	0.677	0.677	0.677	1.22	0.101	0.00	0.00	0.948	1.33	1.10
lgau-1000	NA	3.62	3.77	3.77	3.77	2.89	3.23	3.23	3.23	3.23	3.23	3.23	3.23	1.75	1.08	3.04	3.04	3.04	3.04	3.03
Zkss	NA	3.48	NA	3.7	3.7	3.7	3.7	3.7	3.7	2.07	2.07	2.07	2.07	2.01	2.78	3.48	3.48	3.48	3.48	3.59
CO20pdt	0.808	1.89	2.36	2.75	3.59	3.74	3.33	1.89	3.35	2.75	2.75	2.75	2.75	1.62	1.81	1.78	1.78	1.78	1.78	3.57
CO20pdt.1	1.27	3.83	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.272	0.272	0.272	0.272	0.00	0.00	0.00	0.00	0.32	0.32	3.58
CO20pdt.2	0.546	0.00	0.406	0.00	0.00	0.00	0.00	0.00	0.00	0.31	0.31	0.31	0.31	0.618	0.618	1.04	1.04	0.372	0.372	0.314
CO20pdt.3	0.443	0.00	0.406	0.00	0.00	0.00	0.00	0.00	0.00	0.716	0.716	0.716	0.716	0.00	0.00	0.00	0.00	0.0856	0.0856	0.313
CO20pdt.4	0.474	0.00	0.305	0.32	0.32	0.32	0.32	0.00117	1.13	0.32	0.32	0.32	0.32	0.309	0.309	1.49	1.49	0.00	0.00	0.142
git-sugar	NA	3.16	2.59	2.99	3.57	2.64	3.57	3.57	3.57	3.84	3.84	3.84	3.84	2.56	2.56	3.89	3.89	3.89	3.89	0.449
casaloi...	NA	3.59	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.44	0.44	0.44	0.44	3.52	3.52	3.52	3.52	3.52	3.52	3.85
heikzamber	0.999	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.314	0.314	0.314	0.314	2.12	2.12	0.00	0.00	0.00	0.00	0.449
lavors06-2	NA	1.42	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.307	0.307	0.307	0.307	0.00	0.00	0.00	0.00	0.00	0.00	3.82
lavors06-3	1.42	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.368	0.368	0.368	0.368	0.00	0.00	0.00	0.00	0.00	0.00	0.369
lavors06-4	2.1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.4	0.4	0.4	0.4	0.00	0.00	0.00	0.00	0.00	0.00	0.382
names	0.951	0.845	1.86	0.807	1.86	1.77	2.26	1.86	1.86	0.362	0.362	0.362	0.362	0.00	0.00	0.00	0.00	0.00	0.00	0.308
pept	1.66	3.43	0.639	1.47	1.77	2.26	2.26	1.77	1.99	1.05	1.05	1.05	1.05	1.44	1.44	1.44	1.44	1.31	1.31	0.978
resJ	NA	2.85	1.03	2.89	3.36	1.62	2.02	2.02	2.19	3.02	3.02	3.02	3.02	1.28	1.28	1.96	1.96	0.887	0.887	3.09
res1000	NA	2.49	1.83	1.83	2.69	1.83	3.02	3.02	2.19	2.47	2.47	2.47	2.47	2.16	2.16	2.16	2.16	1.51	1.51	3.11
res2000	NA	1.42	2.26	2.26	2.79	2.26	2.82	2.82	2.3	2.82	2.82	2.82	2.82	2.04	2.04	2.47	2.47	0.824	0.824	3.11
res2kca	NA	3.29	2.07	3.3	3.3	2.16	3.19	3.19	2.3	3.19	3.19	3.19	3.19	3.21	3.21	2.75	2.75	2.94	2.94	2.63
res3000	NA	2.8	1.35	0.807	3.28	2.81	2.26	2.91	2.53	2.91	2.91	2.91	2.91	2.42	2.42	3.19	3.19	2.07	2.07	2.22
res5000	NA	3.39	2.06	3.66	2.1	1.16	1.16	1.16	1.08	2.8	2.8	2.8	2.8	1.08	1.08	1.78	1.78	3.68	3.68	2.63
tiny	0.306	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.334	0.334	0.334	0.334	1.78	1.78	2.55	2.55	1.8	1.8	3.55
water	NA	4.13	2.66	4.17	4.46	4.46	4.9	4.9	3.77	3.86	3.86	3.86	3.86	2.57	2.57	3.82	3.82	4.1	4.1	3.86
Average	1.267	1.828	1.237	1.560	1.939	1.265	2.064	2.064	1.111	2.028	2.028	2.028	2.028	1.259	1.259	1.752	1.752	1.849	1.849	2.027

Table 8: Ψ statistics: blue indicates best.

