

A self-organizing principle for learning nonlinear manifolds

Dimitris K. Agrafiotis* and Huafeng Xu

3-Dimensional Pharmaceuticals, Inc., 665 Stockton Drive, Exton, PA 19341

Edited by Michael Levitt, Stanford University School of Medicine, Stanford, CA, and approved October 9, 2002 (received for review July 17, 2002)

Modern science confronts us with massive amounts of data: expression profiles of thousands of human genes, multimedia documents, subjective judgments on consumer products or political candidates, trade indices, global climate patterns, etc. These data are often highly structured, but that structure is hidden in a complex set of relationships or high-dimensional abstractions. Here we present a self-organizing algorithm for embedding a set of related observations into a low-dimensional space that preserves the intrinsic dimensionality and metric structure of the data. The embedding is carried out by using an iterative pairwise refinement strategy that attempts to preserve local geometry while maintaining a minimum separation between distant objects. In effect, the method views the proximities between remote objects as lower bounds of their true geodesic distances and uses them as a means to impose global structure. Unlike previous approaches, our method can reveal the underlying geometry of the manifold without intensive nearest-neighbor or shortest-path computations and can reproduce the true geodesic distances of the data points in the low-dimensional embedding without requiring that these distances be estimated from the data sample. More importantly, the method is found to scale linearly with the number of points and can be applied to very large data sets that are intractable by conventional embedding procedures.

Extracting the minimum number of independent variables that can describe a set of experimental observations fully is a problem of central importance in science. Most physical processes produce highly correlated outputs, leading to observations that lie on or close to a smooth low-dimensional manifold. Because the dimensionality and nonlinear geometry of that manifold often is embodied in the similarities between the data points, a common approach is to embed the data in a low-dimensional space that best preserves these similarities in the hope that the intrinsic structure of the system will be reflected in the resulting map (1). However, as illustrated in Fig. 1, conventional similarity measures such as the Euclidean distance tend to underestimate the proximity of points on a nonlinear manifold and lead to erroneous embeddings. To remedy this problem, the isometric mapping (ISOMAP) method (2) substitutes an estimated geodesic distance for the conventional Euclidean distance and uses classical multidimensional scaling (MDS) to find the optimum low-dimensional configuration. Although it has been shown that in the limit of infinite training samples, ISOMAP recovers the true dimensionality and geometric structure of the data if it belongs to a certain class of Euclidean manifolds, the proof is of little practical use because the (at least) quadratic complexity of the embedding procedure precludes its use with large data sets. A similar scaling problem plagues locally linear embedding (3), a related approach that produces globally ordered maps by constructing locally linear relationships between the data points. The stochastic proximity embedding (SPE) algorithm described herein utilizes the fact that the geodesic distance is always greater than or equal to the input proximity if the latter is a true metric. Similar to ISOMAP, we assume that the input proximity provides a good approximation to the true geodesic distance when the points are relatively close, which is generally true if the local curvature of

the manifold is not too large. Unlike ISOMAP, however, we circumvent the calculation of approximate geodesic distances between remote points altogether and only require that their distances on the low-dimensional map do not fall below their respective proximities.

Methods

SPE Algorithm. The embedding is carried out by minimizing the stress function,

$$E = \sum_{i < j} \frac{f(d_{ij}, r_{ij})}{r_{ij}} \bigg| \sum_{i < j} r_{ij},$$

where r_{ij} is the input proximity between the i th and j th points, d_{ij} is their Euclidean distance in the low-dimensional space, $f(d_{ij}, r_{ij})$ is the pairwise stress function defined as $f(d_{ij}, r_{ij}) = (d_{ij} - r_{ij})^2$ if $r_{ij} \leq r_c$ or $d_{ij} < r_{ij}$, and $f(d_{ij}, r_{ij}) = 0$ if $r_{ij} > r_c$ and $d_{ij} \geq r_{ij}$, and r_c is the neighborhood radius. The function is minimized by using a self-organizing algorithm that attempts to bring each individual term $f(d_{ij}, r_{ij})$ rapidly to zero. The method starts with an initial configuration and iteratively refines it by repeatedly selecting two points at random and adjusting their coordinates such that their Euclidean distance on the map d_{ij} matches more closely their corresponding proximity r_{ij} . The correction is proportional to the disparity

$$\lambda \frac{|r_{ij} - d_{ij}|}{d_{ij}},$$

where λ is a learning-rate parameter that decreases during the course of the refinement in order to avoid oscillatory behavior. If $r_{ij} > r_c$ and $d_{ij} \geq r_{ij}$, *i.e.*, if the points are nonlocal and their distance on the map is already greater than their proximity r_{ij} , their coordinates remain unchanged. The SPE algorithm proceeds as follows:

1. Initialize the D -dimensional coordinates of the N points, $\{y_{ik}, i = 1, 2, \dots, N, k = 1, 2, \dots, D\}$. Select a cutoff distance r_c .
2. Select a learning rate $\lambda > 0$.
3. Select two points, i and j , at random, retrieve (or evaluate) their proximity in the input space, r_{ij} , and compute their Euclidean distance on the D -dimensional map, $d_{ij} = \|y_i - y_j\|$.
4. If $r_{ij} \leq r_c$, or if $r_{ij} > r_c$ and $d_{ij} < r_{ij}$, update the coordinates y_{ik} and y_{jk} by:

$$y_{ik} \leftarrow y_{ik} + \lambda \frac{1}{2} \frac{r_{ij} - d_{ij}}{d_{ij} + \varepsilon} (y_{ik} - y_{jk}) \text{ and}$$

$$y_{jk} \leftarrow y_{jk} + \lambda \frac{1}{2} \frac{r_{ij} - d_{ij}}{d_{ij} + \varepsilon} (y_{jk} - y_{ik}),$$

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: ISOMAP, isometric mapping; MDS, multidimensional scaling; SPE, stochastic proximity embedding; RMSD, rms deviation.

*To whom correspondence should be addressed. E-mail: agrafiotis@3dp.com.

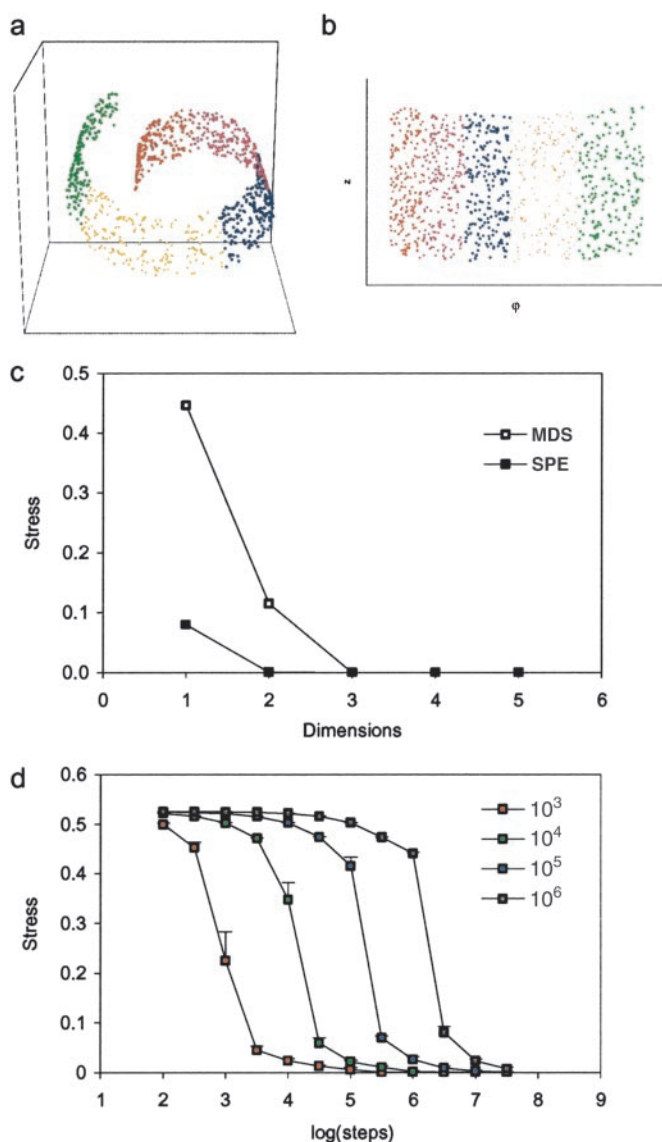


Fig. 1. SPE of the Swiss roll data set described by $\{x = \varphi \cos \varphi, y = \varphi \sin \varphi, z\}$, where φ and z are random numbers in the intervals $[5, 13]$ and $[0, 10]$. The total number of points $N = 1,000$ except for plot *d*. (a) Original data in three-dimensional space. (b) Two-dimensional embedding obtained by SPE. (c) Final stress obtained by SPE (mean and standard deviation over 30 independent runs, each starting from a different initial configuration and random number seed; the standard deviation is too small and therefore barely visible) and MDS as a function of embedding dimensionality. (d) Final stress of two-dimensional embeddings obtained by SPE (mean and standard deviation over 30 independent runs) as a function of simulation length for four data sets containing 10^3 , 10^4 , 10^5 , and 10^6 points. This plot, along with Fig. 3*d*, demonstrates the linear scaling of SPE: a 10-fold increase in sample size results in an ≈ 10 -fold increase in the number of refinement steps that are required to achieve a comparable stress.

where $\varepsilon = 1.0 \times 10^{-10}$ is a small number used to avoid division by zero. If $r_{ij} > r_c$ and $d_{ij} \geq r_{ij}$, leave the coordinates unchanged.

5. Repeat 3 and 4 for a prescribed number of steps, S .
6. Decrease the learning rate λ by a prescribed $\delta\lambda$.
7. Repeat 3–6 for a prescribed number of cycles, C .

In this study, the embeddings were carried out by using 100 refinement cycles, 1 million steps per cycle unless stated otherwise, a linearly decreasing learning rate from 2.0 to 0.1, and a

neighborhood radius at the 10% threshold of all pairwise proximities in the sample as determined by probability sampling. An initial learning rate $\lambda > 1$ was used to induce faster unfolding of the random initial configurations. Alternative learning schedules may also be used.

Data Sets. The data points for the Swiss roll were obtained by generating coordinate triplets $\{x = \varphi \cos \varphi, y = \varphi \sin \varphi, z\}$, where φ and z are random numbers in the intervals $[5, 13]$ and $[0, 10]$, respectively. One thousand conformations of methylpropylether were generated by using a distance geometry algorithm, which uses covalent constraints to establish a set of upper and lower interatomic distance bounds and then attempts to generate conformations that are consistent with these bounds (4). The proximity between conformations was measured by rms deviation (RMSD) (for two conformations, the RMSD is defined as the minimum Euclidean distance between the vectors of atomic coordinates when the two conformations are superimposed through translations and rotations). RMSD is positive, symmetric, and satisfies the triangular inequality[†] and is therefore a valid proximity measure for SPE. The three-component virtual combinatorial library was generated by systematically attaching two aldehyde building blocks to a diamine core according to the reductive amination reaction. Each product was characterized by 117 computed topological indices (5), which subsequently were normalized in the interval $[0, 1]$ and decorrelated by principal component analysis to 26 orthogonal variables that accounted for 99% of the total variance in the data. The Euclidean distance in the resulting 26-dimensional principal component space was used as a proximity measure between two compounds. The principal component analysis preprocessing step was used to eliminate strong linear correlations that are typical of graph-theoretic descriptors and thus accelerate proximity calculations. For the large data sets, the reported stress values were calculated by random sampling of 1 million pairwise distances,

$$E' = \frac{\sum_{(i,j) \in P} f(d_{ij}, r_{ij})}{\sum_{(i,j) \in P} r_{ij}},$$

where $P = \{(i, j) \mid i \neq j, 1 \leq i, j \leq N\}$ consists of randomly sampled pairs. These stochastic stress values have been shown to accurately approximate the true stress (D.K.A., unpublished results).

Results and Discussion

The intrinsic dimensionality of the manifold is revealed by embedding the data in spaces of decreasing dimensions and identifying the point at which the stress effectively vanishes. When applied to the Swiss roll data set (Fig. 1), SPE reliably uncovered the true dimensionality of two. The distances of the points on the two-dimensional map matched the true, analytically derived geodesic distances with a correlation coefficient of 0.9999, indicating a virtually perfect embedding. Similarly, the method was able to detect the intrinsic two-dimensional structure of an ensemble of conformations of methylpropylether compared by RMSD. The coordinate axes on the resulting map correlate very strongly with the molecule's true conformational degrees of freedom, revealing regions of conformational space that are inaccessible because of steric hindrance (Fig. 2).

[†]*Proof:* Suppose that we have three conformations, c_1 , c_2 , and c_3 , and conformations c_1 and c_2 have been superimposed with c_3 such that $\text{RMSD}(c_1, c_3) = \|x_1 - x_3\|$ and $\text{RMSD}(c_2, c_3) = \|x_2 - x_3\|$, where x_i is the vector of atomic coordinates of the i th conformation, and $\|x - y\|$ is the Euclidean distance between vectors x and y . Because RMSD is the minimum Euclidean distance between the vectors of atomic coordinates, we have $\text{RMSD}(c_1, c_2) \leq \|x_1 - x_2\| \leq \|x_1 - x_3\| + \|x_2 - x_3\| = \text{RMSD}(c_1, c_3) + \text{RMSD}(c_2, c_3)$. Q.E.D.

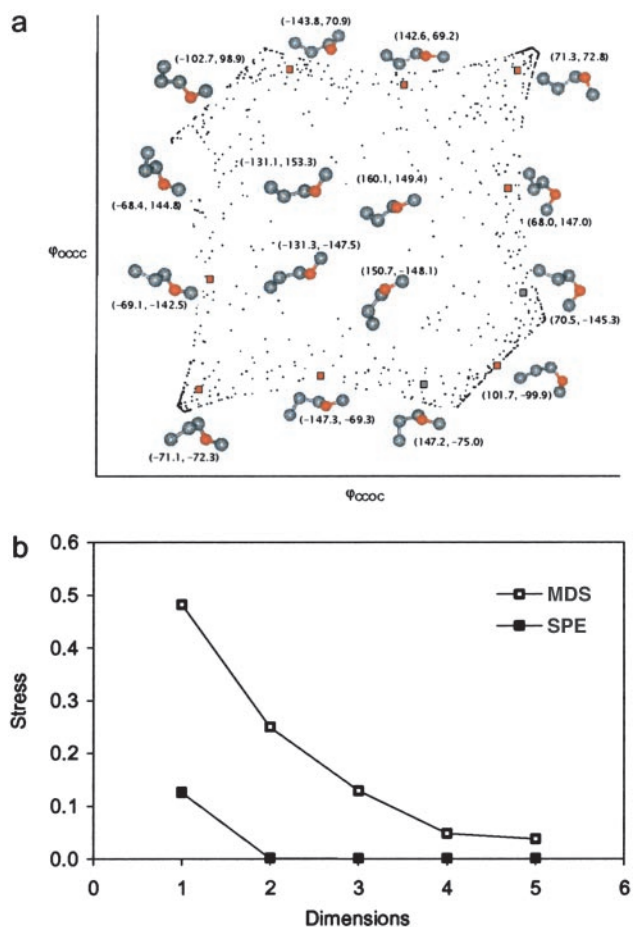


Fig. 2. SPE of 1,000 conformations of methylpropylether, $C_1C_2C_3O_4C_5$, generated by a distance geometry algorithm and compared by RMSD. (a) Two-dimensional embedding obtained by SPE. Representative conformations are shown next to highlighted points in different parts of the map along with the corresponding torsional angles, $\varphi_{C_2C_3O_4C_5}$ and $\varphi_{C_1C_2C_3O_4}$, in parentheses. The horizontal and vertical directions represent rotation around the C_3-O_4 and C_2-C_3 bonds, respectively. The unoccupied upper-left and bottom-right corners represent conformations that are inaccessible because of the steric hindrance between the two terminal carbon atoms C_1 and C_5 . (b) Final stress obtained by SPE (mean and standard deviation over 30 independent runs) and MDS as a function of embedding dimensionality.

SPE can also produce meaningful low-dimensional representations of more complex data sets that do not have a clear manifold geometry. The embedding of a combinatorial chemical library illustrated in Fig. 3 shows that the method is able to preserve local neighborhoods of closely related compounds while maintaining a chemically meaningful global structure (Fig. 3). Although the intrinsic dimensionality of this data set is substantially higher than two, the two-dimensional map exhibits global order and continuity, as manifested by the dominant role of molecular weight, and the presence of variation patterns that correspond to chemically distinguishing features such as chain length, ring structure, and halogen content (6).

In contrast to previous stochastic approaches of nonlinear manifold learning that preferentially preserve local over global distances, SPE follows the philosophy, first proposed and exploited in the ISOMAP method, that the geodesic distances characterize the topological structure better than the apparent Euclidean distances. Unlike previous approaches that assign arbitrary weighting factors to the local and remote distances (7) or utterly disregard the remote distances (8, 9), SPE differen-

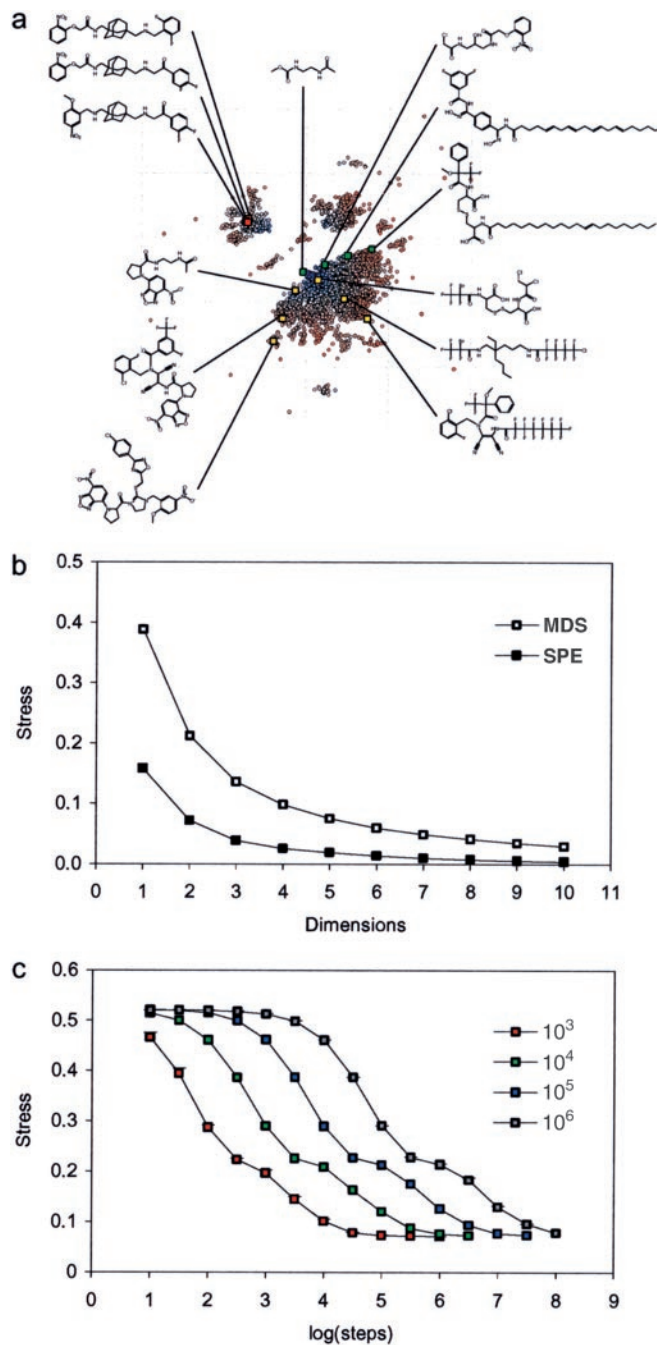


Fig. 3. SPE of the diamine combinatorial library. The total number of points $N = 1,000$ except for plot c. (a) Two-dimensional embedding obtained by SPE. (b) Final stress obtained by SPE (mean and standard deviation over 30 independent runs) and MDS as a function of embedding dimensionality. (c) Final stress of two-dimensional embeddings obtained by SPE (mean and standard deviation over 30 independent runs) as a function of simulation length for four data sets containing 10^3 , 10^4 , 10^5 , and 10^6 compounds.

tiates local from remote distances by their intrinsic relationship to the true geodesic distances and utilizes both types accordingly. Therefore, it can preserve the local geometry and the global topology of the manifold better than previous approaches.

Although SPE does not offer the global optimality guarantees of ISOMAP or locally linear embedding, it works very well in practice. As illustrated by the variances in Figs. 1c and 2b, the method converges reliably to the global minimum when the data

are embedded in a space of the intrinsic dimensionality (and to a low-stress configuration in fewer dimensions) regardless of the starting configuration and initialization conditions. More importantly, when applied to data sets of increasing size drawn from the same probability distribution (and therefore expected to have comparable stress), the number of sampling steps required to reach a particular stress increases in linear fashion (Figs. 1d and 3c). The memory requirements of the algorithm grow linearly as well, because the proximities can be computed on demand and need not be stored explicitly.

These characteristics are attributed to the stochastic nature of the refinement scheme and the vast redundancy of the distance matrix. SPE is reminiscent of the stochastic approximation approach introduced by Robbins and Monroe (10) and popularized by the back-propagation algorithm of Rumelhart *et al.* (11). The direction of each pairwise refinement can be thought of as an instantaneous gradient, a stochastic approximation of the true gradient of the stress function. For sufficiently small numbers of λ , the average direction of these refinements approximates the direction of steepest descent (for a mathematical justification of the above statement, see *Supporting Text*, which is published as supporting information on the PNAS web site, www.pnas.org). Unlike classical gradient minimization schemes, the use of stochastic gradients changes the effective error function in each step, and the method becomes less susceptible to local minima. In addition, the method exploits the redundancy in the interpoint distances through probability sampling. Indeed, the relative configuration of $N > D$ points in a D -dimensional space can be described fully by using only $(N - D/2 - 1)/(D + 1)$ distances, which is consistent with the linear complexity of SPE. Linear scaling in both time and memory is critical in modern data mining where large data sets abound.

Of course, just like ISOMAP and locally linear embedding, SPE depends on the choice of the neighborhood radius, r_c . If r_c is too large, the local neighborhoods will include data points from other branches of the manifold, shortcutting them and leading to substantial errors in the final embedding. If it is too small, it will lead to discontinuities, causing the manifold to fragment into a large number of disconnected clusters. The ideal threshold can be determined by examining the stability of the algorithm over a range of neighborhood radii as prescribed by Tenenbaum (12). By setting r_c to infinity, SPE can produce nonlinear maps that are essentially identical to those derived by classical MDS (D.K.A., unpublished results). In this case, the efficiency of the algorithm is even more impressive, because virtually all the randomly chosen pairs result in “productive” work. When a cutoff is used, once the general structure of the

map has been established, the majority of pairwise comparisons do not result in any refinement, because most of the remote points already are separated beyond their lower bounds. This situation can be improved by caching and resampling neighbors during the course of the refinement (S. Izrailev, H.X., and D.K.A., unpublished results).

One potential criticism of SPE is of its numerous free parameters. Superficially, in addition to the neighborhood radius r_c , SPE also depends on the number of cycles C , the number of steps per cycle S , the initial learning rate λ_0 , the annealing schedule for the learning rate, and the initial configuration. In practice, however, we find that the final results are generally insensitive to these choices (see *Supporting Text*). For most applications, we recommend $C = 100$, $\lambda_0 = 2.0$, $\delta\lambda = (\lambda_0 - \lambda_1)/(C - 1)$, where λ_1 is the final learning rate that can be any arbitrary small number, typically between 0.01 and 0.1. The initial configuration can be any random distribution of the N points in a D -dimensional hypercube of side length $N^{1/D}r_c$. The number of steps per cycle S , or equivalently the total number of pairwise adjustments, therefore remains as the *only* extra free parameter besides the neighborhood radius. S should be increased linearly with N as SPE’s empirical characteristic of linear scaling suggests. Moreover, we find that the number of steps should also be increased for structures with large curvatures. An obvious procedure is to use increasing S until the final stress no longer diminishes.

SPE is general and can be applied to any problem for which nonlinearity complicates the use of conventional methods such as principal component analysis and MDS, and a sensible proximity measure, similar to the ones mentioned above, can be defined. The method is trivial to implement and computationally inexpensive and can be used as a tool for exploratory data analysis and visualization. The coordinates produced by SPE can be used further as input to a parametric learner in order to derive an explicit mapping function between the observation and embedded spaces (13). Finally, because it fundamentally seeks an embedding that is consistent with a set of upper and lower distance bounds (the proximity of neighboring points can be viewed as a degenerate distance range with identical lower and upper bounds), SPE also can be applied to an important class of distance geometry problems including conformational analysis (14), NMR structure determination, and protein-structure prediction (15) (H.X., S. Izrailev, and D.K.A., unpublished results).

We thank Dr. F. Raymond Salemme, Victor S. Lobanov, Sergei Izrailev, and Dmitrii N. Rassokhin of 3-Dimensional Pharmaceuticals for many useful discussions.

- Borg, I. & Groenen, P. J. F. (1997) *Modern Multidimensional Scaling: Theory and Applications* (Springer, New York).
- Tenenbaum, J. B., de Silva, V. & Langford, J. C. (2000) *Science* **290**, 2319–2323.
- Roweis, S. T. & Saul, L. K. (2000) *Science* **290**, 2323–2326.
- Crippen, G. M. & Havel, T. F. (1988) *Distance Geometry and Molecular Conformation* (Research Studies, Somerset, U.K.).
- Kier, L. B. & Hall, L. H. (1986) *Molecular Connectivity in Structure-Activity Analysis* (Wiley, New York).
- Agrafiotis, D. K., Lobanov, V. S. & Salemme, F. R. (2002) *Nat. Rev. Drug Discov.* **1**, 337–346.
- Shepard, R. N. & Carroll, J. D. (1965) in *International Symposium on Multivariate Analysis*, ed. Krishnaiah, P. R. (Academic, New York), pp. 561–592.
- Demartines, P. & Héroult, J. (1997) *IEEE Trans. Neural Netw.* **8**, 148–154.
- Héroult, J., Jausions-Picaud, C. & Guérin-Dugué, A. (1999) *Int. Work Conf. Artif. Nat. Neural Netw.* **2**, 625–634.
- Robbins, H. & Monroe, S. (1951) *Ann. Math. Stat.* **22**, 400–407.
- Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986) *Nature* **323**, 533–536.
- Tenenbaum, J. B. (2002) *Science* **295**, 7a.
- Agrafiotis, D. K. & Lobanov, V. S. (2000) *J. Chem. Inf. Comput. Sci.* **40**, 1356–1362.
- Spellmeyer, D. C., Wong, A. K., Bower, M. J. & Blaney, J. M. (1997) *J. Mol. Graph. Model.* **15**, 18–36.
- Havel, T. F. & Wüthrich, K. (1985) *J. Mol. Biol.* **182**, 281–294.