A Geodesic Framework for Analyzing Molecular Similarities

Dimitris K. Agrafiotis* and Huafeng Xu

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

Received October 31, 2002

A fast self-organizing algorithm for extracting the minimum number of independent variables that can fully describe a set of observations was recently described (Agrafiotis, D. K.; Xu, H. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 15869−15872). The method, called stochastic proximity embedding (SPE), attempts to generate low-dimensional Euclidean maps that best preserve the similarities between a set of related objects. Unlike conventional multidimensional scaling (MDS) and nonlinear mapping (NLM), SPE preserves only local relationships and, by doing so, reveals the intrinsic dimensionality and metric structure of the data. Its success depends critically on the choice of the neighborhood radius, which should be consistent with the local curvature of the underlying manifold. Here, we describe a procedure for determining that radius by examining the tradeoff between the stress function and the number of connected components in the neighborhood graph and show that it can be used to produce meaningful maps in any embedding dimension. The power of the algorithm is illustrated in two major areas of computational drug design: conformational analysis and diversity profiling of large chemical libraries.

I. INTRODUCTION

Virtually all marketed drugs result from the optimization of a lead compound identified through random screening or serendipitous observation of a pharmaceutically relevant side effect. This optimization process involves systematic modification of the initial lead driven by structure−activity data, synthetic feasibility, and chemical intuition. Central to this approach is the thesis that chemically similar compounds tend to exhibit similar biological properties. Although medicinal chemists can cope with the inherent vagueness of this concept, computer-aided methods necessitate a more rigorous definition.

Molecular similarity is typically quantified by a numerical index derived either through direct comparison or through the measurement of a set of characteristic features, which are combined using a suitable distance measure.[1] Since it is not possible to know a priori which molecular properties are most relevant to the problem at hand, a comprehensive set of descriptors is usually employed, chosen based on experience, software availability, and computational cost. This approach, however, is susceptible to the *curse of dimensionality*[2]—high-dimensional spaces are sparse and counterintuitive, and their structure cannot be easily extracted with conventional graphical techniques. Fortunately, most multivariate data in R^d are not truly d-dimensional. Strong correlations between the input variables leads to patterns that lie on or close to a smooth lower-dimensional manifold. Extracting the intrinsic dimensionality and metric structure of that manifold is a problem of paramount importance if the data are to be properly modeled.

High-dimensional spaces are encountered in many important areas of computational drug design, including similarity searching, diversity and drug-like profiling of large chemical libraries[3] and quantitative structure−activity modeling.[4] The classical methods for reducing the dimensionality of such spaces are principal component analysis (PCA)[5] and multidimensional scaling (MDS).[6] The former reduces a set of partially cross-correlated data into a smaller set of orthogonal variables with minimal loss in the contribution to variation, whereas the latter produces an embedding that preserves the interpoint distances. Although these methods work well with linear or quasi-linear subspaces, they fail to detect nonlinear structures, curved manifolds, and arbitrarily shaped clusters.

The primary failure of MDS lies in the fact that it tries to preserve all pairwise distances in the data sample, both local and remote. Indeed, given a set of N objects, a symmetric matrix, r_{ij}, of relationships (proximities) between these objects, and a set of images on a D-dimensional display map $\{\boldsymbol{x}_i, i = 1, 2,.., N; \boldsymbol{x}_i \in R^D\}$, MDS attempts to place $\boldsymbol{x}_i$ onto the plane in such a way that their Euclidean distances $d_{ij} = ||\boldsymbol{x}_i - \boldsymbol{x}_j||$ approximate as closely as possible the corresponding values r_{ij}. This is typically accomplished by minimizing an error function that measures the discrepancy between the input and output distances, such as Kruskal's stress:

$$S = \sqrt{\sum_{i<j}(d_{ij} - r_{ij})^2 \Big/ \sum_{i<j} d_{ij}^2}$$

However, it has been known for a long time that conventional similarity measures such as the Euclidean distance tend to underestimate the proximity of points on a nonlinear manifold and lead to erroneous embeddings.[7,8] Sammon's nonlinear mapping (NLM) algorithm[9] partly alleviates this problem by introducing a normalization factor in the error function to give increasing weight to short-range distances over long-range ones:

$$S = \sum_{i<j} \frac{(d_{ij} - r_{ij})^2}{r_{ij}} \Big/ \sum_{i<j} r_{ij}$$

* Corresponding author phone: (610)458-6045; fax: (610)458-8249; e-mail: dimitris.agrafiotis@3dp.com.

10.1021/ci025631m CCC: $25.00
Published on Web 02/01/2003

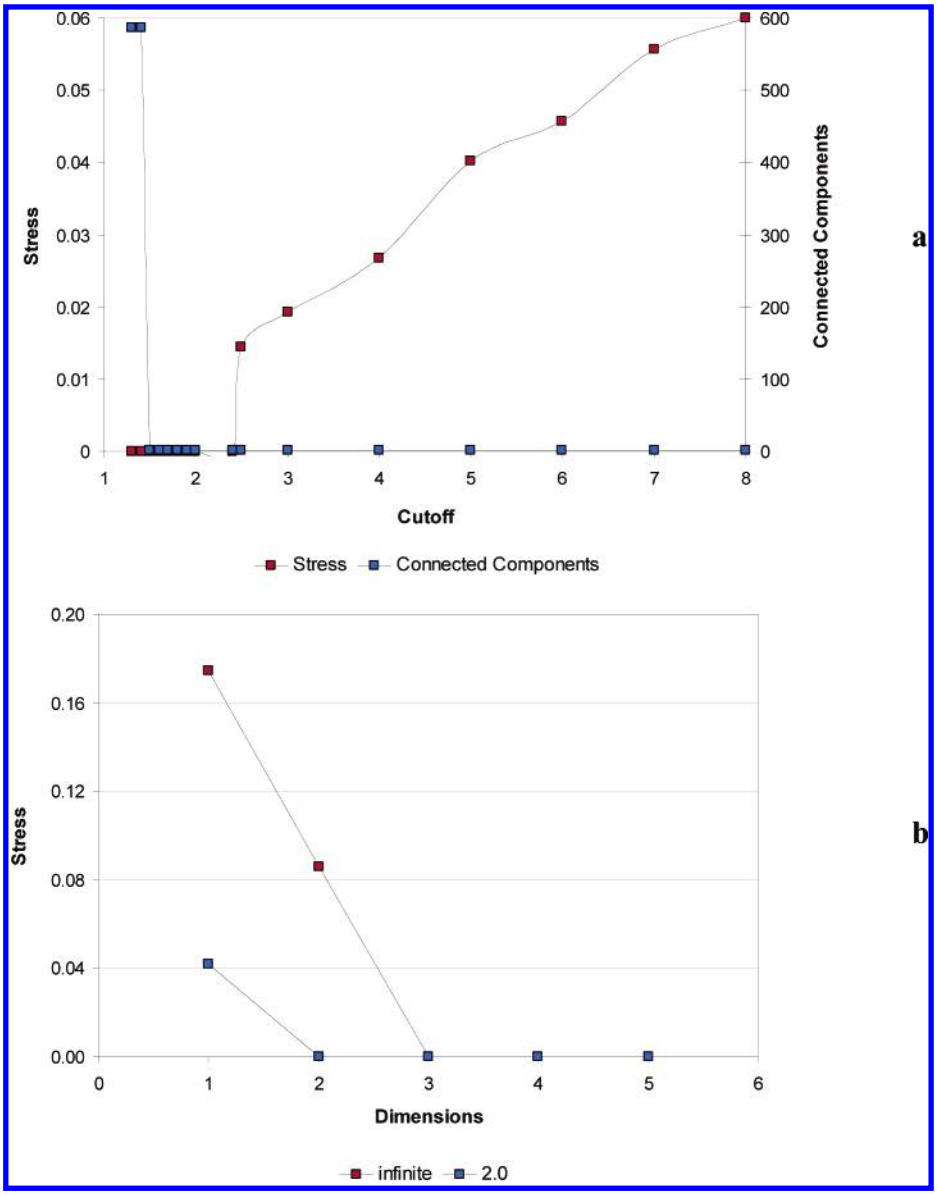


Figure 1. Effect of cutoff on the stress, fragmentation, and dimensionality of the nanotube data set. **(**a) Stress and number of connected components of the two-dimensional SPE map as a function of cutoff. Any value between 1.41 Å and 2.44 Å will produce the correct embedding illustrated in Figure 2c. (b) Final stress of SPE maps as a function of the embedding dimension with (blue, $r_c = 2$) and without a cutoff (red, $r_c = \infty$). Without a cutoff, the stress vanishes in three dimensions, reproducing the *apparent* dimensionality of the data sample. With a proper cutoff, the stress vanishes in two dimensions, reproducing the *intrinsic* dimensionality.

This scheme, however, is arbitrary and fails with highly folded topologies. To remedy this problem, Tenenbaum et al.[10] introduced the ISOMAP method, which uses an estimated geodesic distance instead of the conventional Euclidean one as input to the MDS procedure. The geodesic distances are estimated by connecting each point to its nearest neighbors and then tracing the shortest paths between all pairs of points on the resulting graph. Although it was 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 since the (at least) quadratic complexity of the embedding algorithm precludes its use with large data sets. A similar scaling problem plagues locally linear embedding (LLE),[11] a related approach that produces globally ordered maps by constructing locally linear relationships between the data points.

Recently, we introduced stochastic proximity embedding (SPE),[12] a novel self-organizing scheme that addresses the key limitations of ISOMAP and LLE. SPE builds on the same geodesic principle first proposed and exploited by ISOMAP but introduces two important algorithmic advances: (1) it circumvents the calculation of estimated geodesic distances, and (2) it uses a pairwise refinement scheme that does not require the complete distance (d_{ij}) or proximity (r_{ij}) matrix and scales linearly with the number of points. The method minimizes the stress function

$$S = \sum_{i<j} \frac{f(d_{ij}, r_{ij})}{r_{ij}} \Big/ \sum_{i<j} r_{ij}$$

where $f(d_{ij}, r_{ij})$ is the pairwise stress defined as $f(d_{ij}, r_{ij}) = (d_{ij} - r_{ij})^2$ if $r_{ij} \le r_c$ or $d_{ij} < r_{ij}$, and $f(d_{ij}, r_{ij}) = 0$ if $r_{ij} > r_c$ and $d_{ij} \ge r_{ij}$, and r_c is a predefined neighborhood radius.

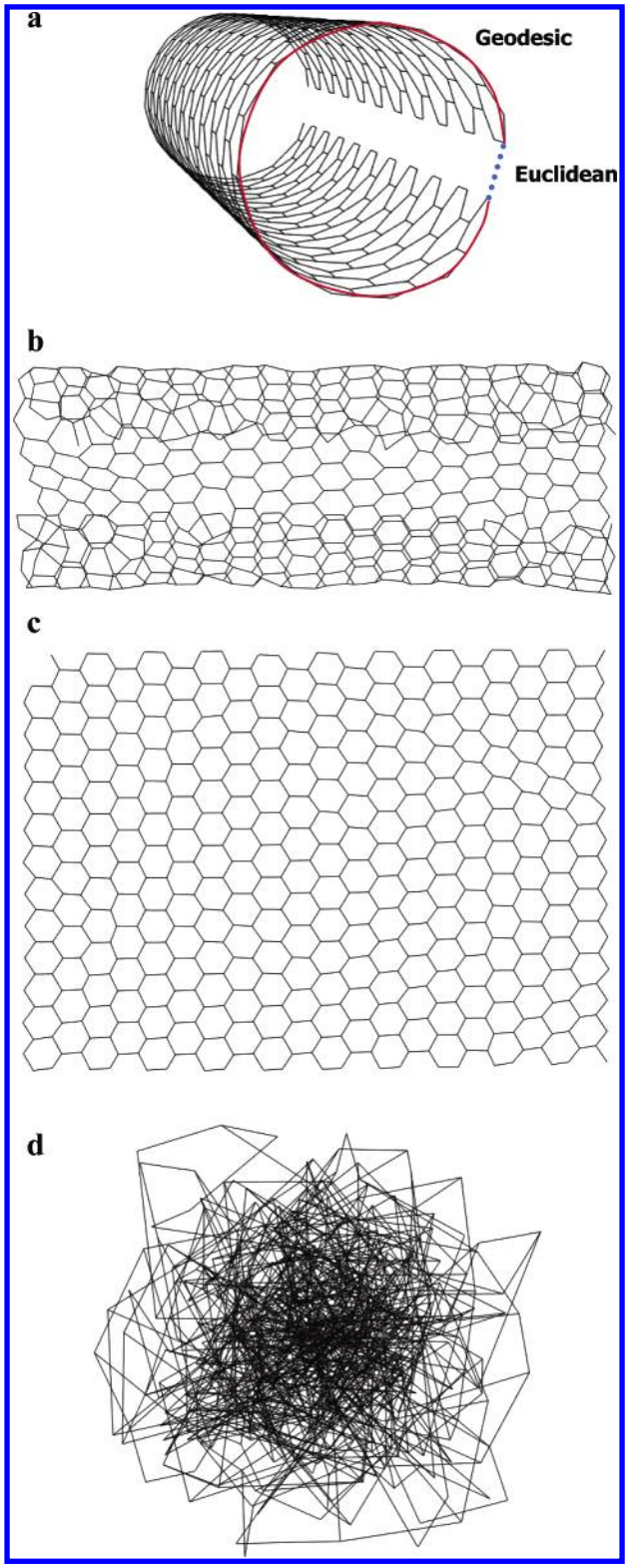


Figure 2. Two-dimensional embeddings of the nanotube obtained by SPE. (a) Original three-dimensional structure. (b) Two-dimensional map obtained with $r_c = \infty$. (c) Two-dimensional map obtained with $r_c = 2.0$. (d) Two-dimensional map obtained with $r_c = 1.0$. The original connectivity of the atoms is preserved in all of these maps.

This is accomplished using a stochastic approximation of steepest descent 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 so that their Euclidean distance on the map d_{ij} matches more closely their corresponding proximity r_{ij}. The correction is proportional to the disparity $\lambda^{|r_{ij} - d_{ij}|/d_{ij}}$, 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. Unlike conventional MDS, SPE preserves exact distances between neighboring points and lower bounds between remote points, thus allowing the manifold to unfold and reveal its true intrinsic dimensionality. In essence, the method views the input distances between remote points as lower bounds of their true geodesic distances and uses them as a means to impose global structure.

Just like ISOMAP and LLE, SPE depends critically 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. In this work, we present a method for determining a reasonable cutoff by examining the tradeoff between the stress function and the number of connected components in the neighborhood graph and show that it can be used to produce meaningful maps in any embedding dimension. This method introduces a new way of thinking about chemical neighborhoods, rooted in modeling the nonlinear geometry of chemical space.

II. METHODS

Stochastic Proximity Embedding. The SPE algorithm proceeds as follows:

1. Initialize the D-dimensional coordinates of the N points, $\{x_{ik}; i = 1,2,...,N; k = 1,2,...,D\}$. Select a cutoff distance r_c and an initial learning rate $\lambda > 0$.

2. 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, $d_{ij} = ||\boldsymbol{x}_i - \boldsymbol{x}_j||$. If $r_{ij} \leq r_c$, or if $r_{ij} > r_c$ and $d_{ij} < r_{ij}$, update the coordinates $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ by

$$x_i \leftarrow x_i + \lambda \frac{1}{2} \frac{r_{ij} - d_{ij}}{d_{ij} + \epsilon}(x_i - x_j)$$

and

$$x_j \leftarrow x_j + \lambda \frac{1}{2} \frac{r_{ij} - d_{ij}}{d_{ij} + \epsilon}(x_j - x_i)$$

where ϵ is a small number used to avoid division by zero (here set to 1.0×10^{-10}). If $r_{ij} > r_c$ and $d_{ij} \geq r_{ij}$, leave the coordinates unchanged.

3. Repeat (2) for a prescribed number of steps, S.

4. Decrease the learning rate λ by a prescribed $\delta\lambda$.

5. Repeat (2)−(4) for a prescribed number of cycles, C.

Connected Components. The ideal cutoff is determined by examining the tradeoff between the stress function and the number of connected components in the neighborhood graph at different values of r_c. For a given value of r_c, the neighborhood graph is an undirected graph that contains a

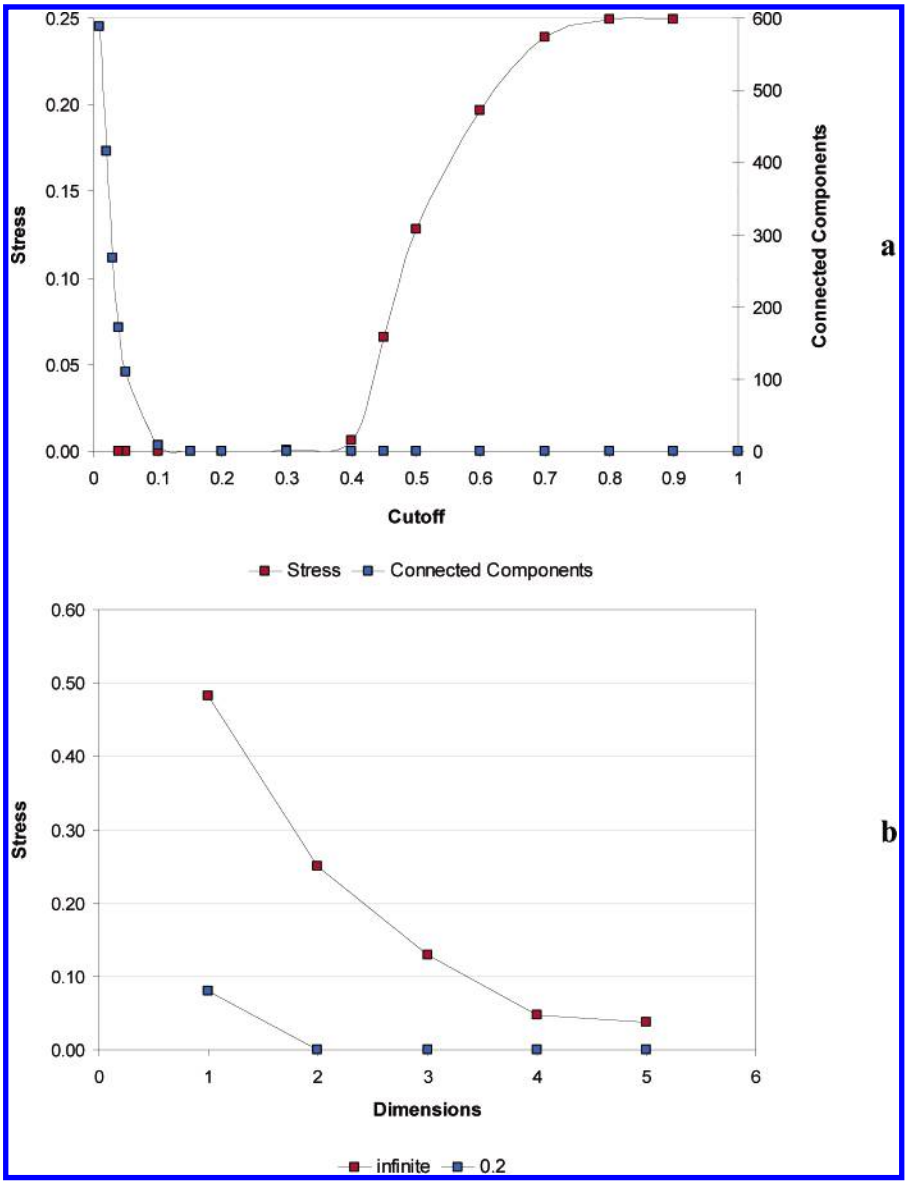


Figure 3. Effect of cutoff on the stress, fragmentation, and dimensionality of the methylpropyl ether data set. (a) Stress and number of connected components of the two-dimensional SPE map as a function of cutoff. Any value between 0.1 Å and 4.0 Å will produce the correct embedding illustrated in Figure 4c. (b) Final stress of SPE map as a function of the embedding dimension with (blue, $r_c = 0.2$) and without a cutoff (red, $r_c = \infty$). As in Figure 1b, the intrinsic dimensionality of the data set is recovered only when a proper cutoff is used.

vertex for every point in the data set and an edge between any pair of points whose proximity is less than or equal to r_c. Connected components represent distinct fragments of the graph—two vertices are said to belong to the same component if there is a path between them. Efficient algorithms for computing connected components can be found in ref 13. Since the cutoff is dependent on the intrinsic curvature and sampling frequency of the manifold, the procedure outlined above should be relatively insensitive to the embedding dimension.

Data Sets. The algorithm was tested on three different chemical data sets.

(1) Nanotube. The first data set is the three-dimensional structure of a nanotube comprised of 565 carbon atoms. Since SPE is not applicable to non-Euclidean manifolds such as cylinders, tori, etc., the nanotube was "cut" by removing a row of atoms from the original structure file. The resulting data set is illustrated in Figure 2a.

(2) Methylpropyl Ether. The second data set consists of 1000 conformations of methylpropyl ether (MPE), generated using a variant of SPE for conformational sampling.[14] Just like conventional distance geometry,[15] this method 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. The embedding is carried out by minimizing the error function:

$$S = S_d + S_v = \sum_{i<j} f(d_{ij}, l_{ij}, u_{ij}) + \alpha \sum_k h(V_k, V_k^l, V_k^u)$$

The first term gives the violation of the distance constraints, where $f(d_{ij}, l_{ij}, u_{ij}) = (d_{ij}^2 - l_{ij}^2/l_{ij}^2)^2$ if $d_{ij} < l_{ij}$, $f(d_{ij}, l_{ij}, u_{ij}) = (d_{ij}^2 - u_{ij}^2/u_{ij}^2)^2$ if $d_{ij} > u_{ij}$, and $f(d_{ij}, l_{ij}, u_{ij}) = 0$ otherwise, the second term gives the violation of the volume constraints, where $h(V_k, V_k^l, V_k^u) = (V_k - V_k^l)^2$ if $V_k < V_k^l$, $h(V_k, V_k^l, V_k^u) = (V_k - V_k^u)^2$ if $V_k > V_k^u$, and $h(V_k, V_k^l, V_k^u) = 0$ otherwise, and α is a scaling factor (here $\alpha = 0.1$). Volume constraints

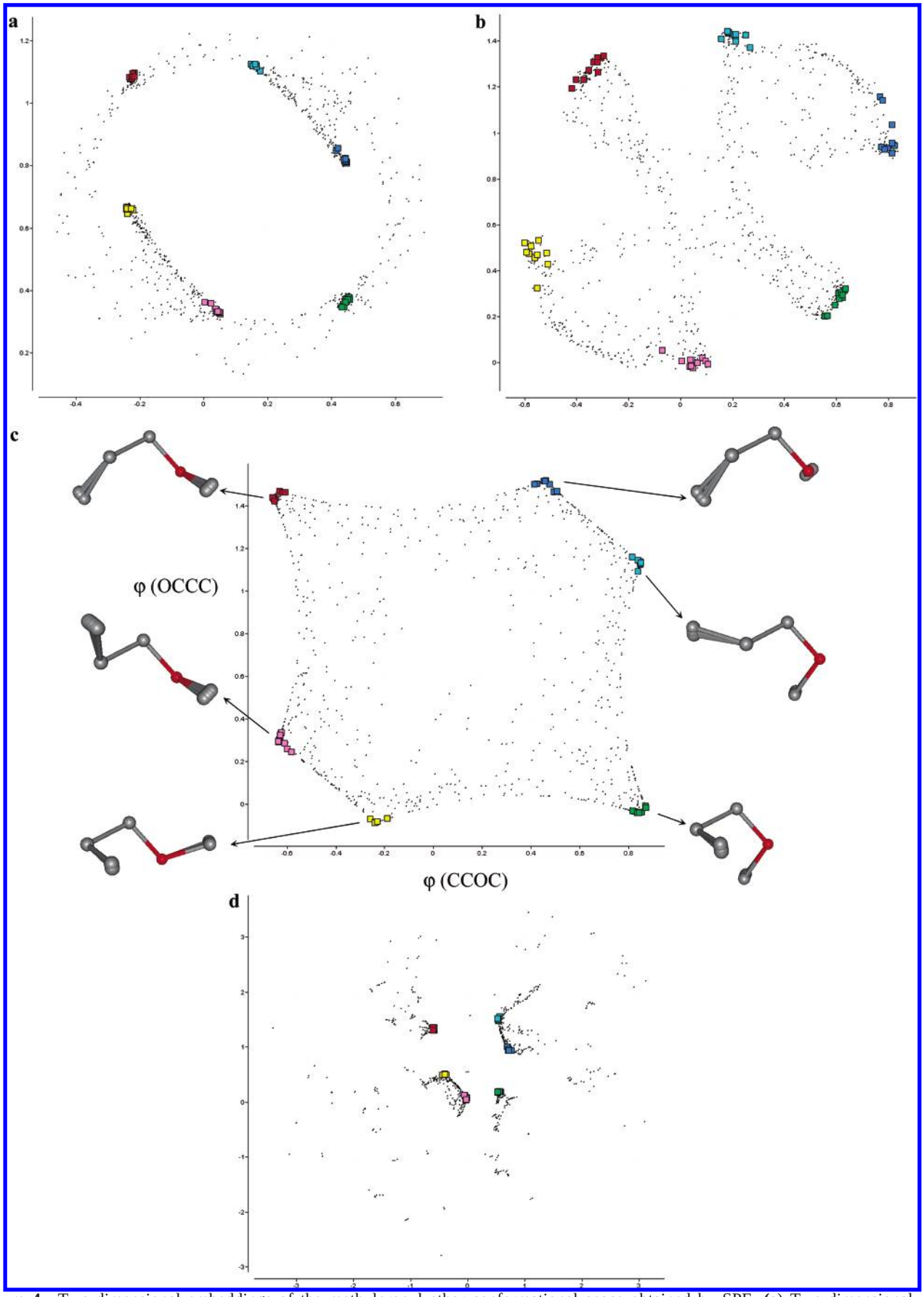


Figure 4. Two-dimensional embeddings of the methylpropyl ether conformational space obtained by SPE. **(**a) Two-dimensional map obtained with $r_c = \infty$. (b) Two-dimensional map obtained with $r_c = 0.5$. (c) Two-dimensional map obtained with $r_c = 0.2$. Representative conformations are shown next to the highlighted points, all superimposed along the central CCO atoms and displayed in the same orientation. The same points are highlighted with like colors in panels a, b, and d. The horizontal and vertical directions correspond to rotations around the central CO and CC bonds, respectively. (d) Two-dimensional map obtained with $r_c = 0.05$.

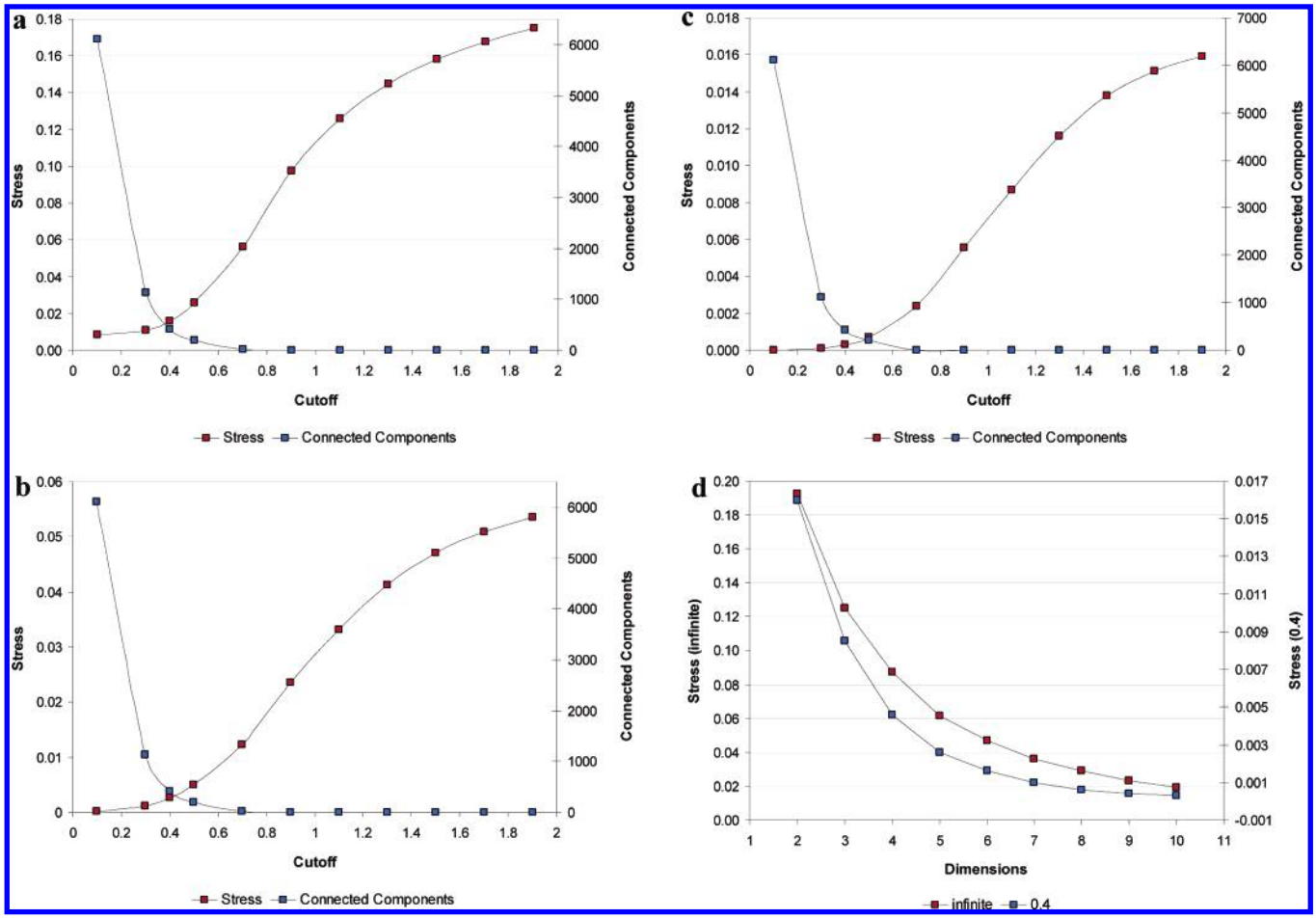


Figure 5. Effect of cutoff on the stress, fragmentation, and dimensionality of the amination library. (a) Stress and number of connected components of the two-dimensional SPE map as a function of cutoff. (b) Stress and number of connected components of the five-dimensional SPE map as a function of cutoff. (c) Stress and number of connected components of the 10-dimensional SPE map as a function of cutoff. (d) Final stress of SPE map as a function of the embedding dimension with (blue, $r_c = 0.4$) and without a cutoff (red, $r_c = \infty$). For clarity, these curves are shown on a different scale.

prevent the signed volume V_{ijkl} formed by four atoms i,j,k,l from exceeding certain limits and are used to enforce planarity of conjugate systems and correct chirality of stereocenters (not applicable in the case of MPE). Minimizing the error function S with respect to the atomic coordinates generates conformations that satisfy the distance and volume constraints. Similar to the procedure described above, the error is minimized by randomly selecting a distance or volume constraint k and moving the positions of the atoms involved in the direction that minimizes the individual error $f(d_{ij}, l_{ij}, u_{ij})$ or $h(V_k, V_k^l, V_k^u)$, respectively. The proximity between conformations was measured by the RMSD, which is defined as the minimum Euclidean distance between the atomic coordinate vectors of two conformations superimposed through translations and rotations.

(3) Amination Library. The third data set represents a two-component virtual combinatorial library[3] containing 10 000 compounds derived by combining 100 amines and 100 aldehydes using the reductive amination reaction. Each of the products was described by 117 topological descriptors including molecular connectivity indices, kappa shape indices, subgraph counts, information-theoretic indices, Bonchev-Trinajstić indices, and topological state indices.[16] To eliminate strong linear correlations, which are typical of graph-theoretic descriptors, the data were normalized and decorrelated using PCA. Molecular dissimilarity was defined as the Euclidean distance in the latent variable space formed by the 23 principal components that accounted for 99% of the total variance in the data.

Implementation. All programs were implemented in the C++ programming language and are part of the DirectedDiversity software suite.[17] All calculations were carried out on a Dell Inspiron 8000 laptop computer equipped with a 1.0 GHz Pentium III Intel processor running Windows 2000 Professional.

III. RESULTS AND DISCUSSION

The first example illustrates the principle of manifold learning in situations where the intrinsic dimensionality of the system is known from simple geometrical considerations. Consider the nanotube illustrated in Figure 2a. Although the object is described in three dimensions, its intrinsic dimensionality is 2 since all the atoms lie on a surface and their location can be fully described by two variables—the angle ϑ and the depth z. A useful feature of this data set is that the distribution of atoms on the nanotube is nonrandom. The connectivity of the molecule imposes strict interatomic distance constraints—1.41 Å for 1−2 (bonded) atoms, 2.44 Å for 1−3 atoms, and 2.84 Å for 1−4 atoms—resulting in a well-defined pattern of unfolding (Figure 1a). SPE produces the correct embedding (Figure 2c) at any distance larger than

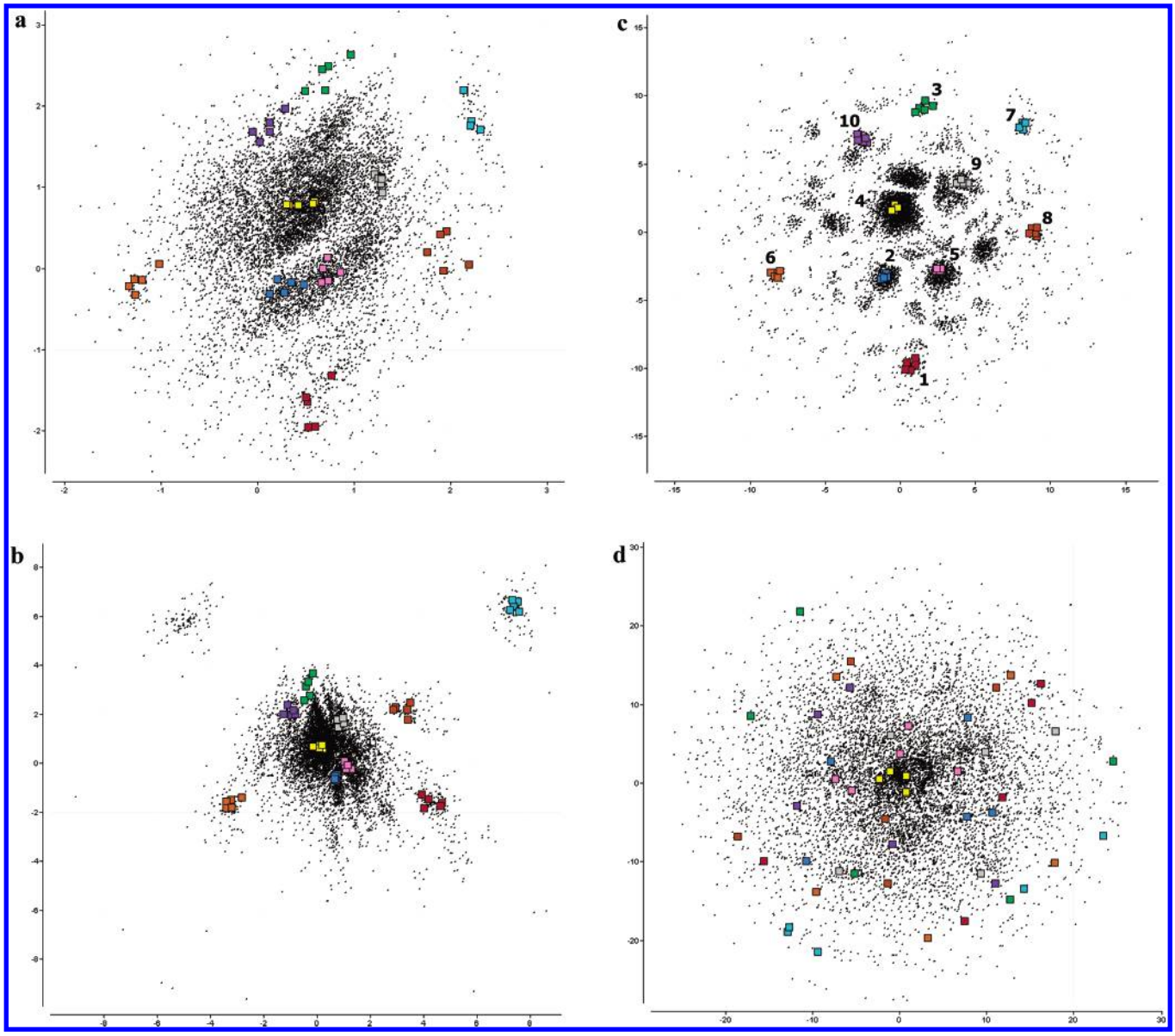
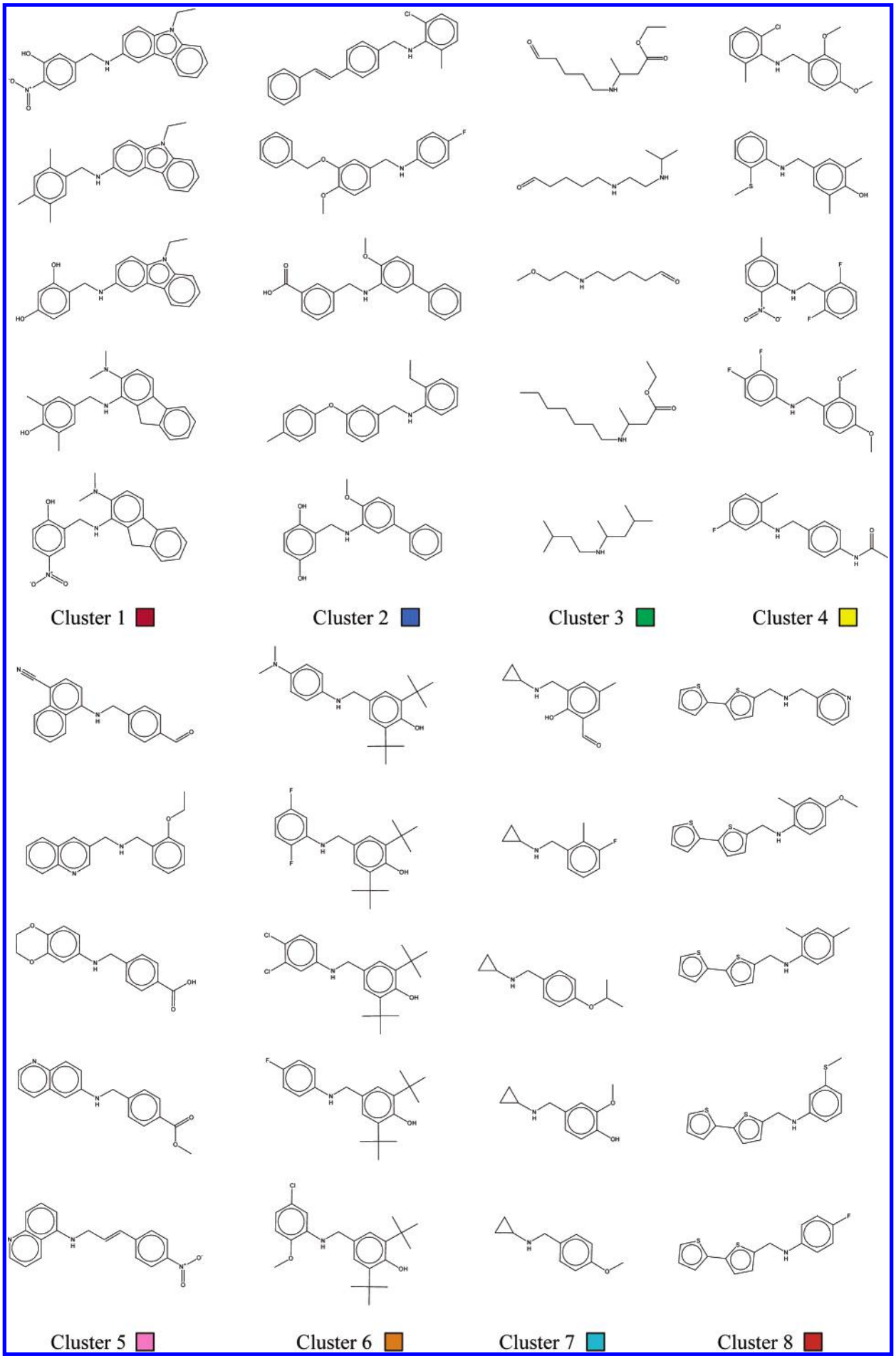


Figure 6. Two-dimensional embeddings of the amination library obtained by SPE. (a) Two-dimensional map obtained with $r_c = \infty$. (b) Two-dimensional map obtained with $r_c = 0.9$. (c) Two-dimensional map obtained with $r_c = 0.4$. (d) Two-dimensional map obtained with $r_c = 0.1$.

1.41 Å and smaller than 2.44 Å. Below 1.41 Å, all connectivity information is lost, and the atoms are placed essentially randomly on the nonlinear map (Figure 2d). This is manifested in zero stress (no distances need to be preserved) but a sharp increase in the number of connected components (Figure 1a). When the cutoff exceeds 2.44 Å, the atoms at the two opposite ends of the surface where the nanotube was cut are short-circuited, causing the surface to fold and conceal its intrinsic dimension (Figure 2b). Since the cutoff reflects the sampling density and local curvature of the manifold and is independent of the embedding dimension, the intrinsic dimension is properly reflected in the scree plot in Figure 1b.

A far more interesting mapping is that of the conformational space of methylpropyl ether. This example illustrates two distinct applications of SPE—one for constructing the actual conformations and one for visualizing them in a low-dimensional Euclidean space. Since RMSD is based on superposition of atomic coordinates, the apparent dimensionality of that space is 9 ($5 \times 3-6$; see Figure 3b). However, it is known from chemical intuition that MPE possesses only two conformational degrees of freedom—the rotations around the central CC and CO bonds. As illustrated in Figure 3, this structure is visible to SPE only when the neighborhood size is between 0.1 and 0.4 Å. In this case, the coordinate axes on the resulting map correlate very strongly with the molecule's intrinsic conformational degrees of freedom (the two rotatable bonds) and reveal regions of conformational space that are inaccessible due to steric hindrance (the missing upper-right and bottom-left triangles in Figure 4c). Larger cutoff values force SPE to bring close together unrelated conformations, while smaller ones cause it to fragment into a large number of conformational "islands". In this case, the notion of the nonlinear manifold has a direct physical interpretation—the geodesic distance between two conformations corresponds to the trajectory that must be followed to convert one to the other. RMSD underestimates the length of this trajectory because it assumes


Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8

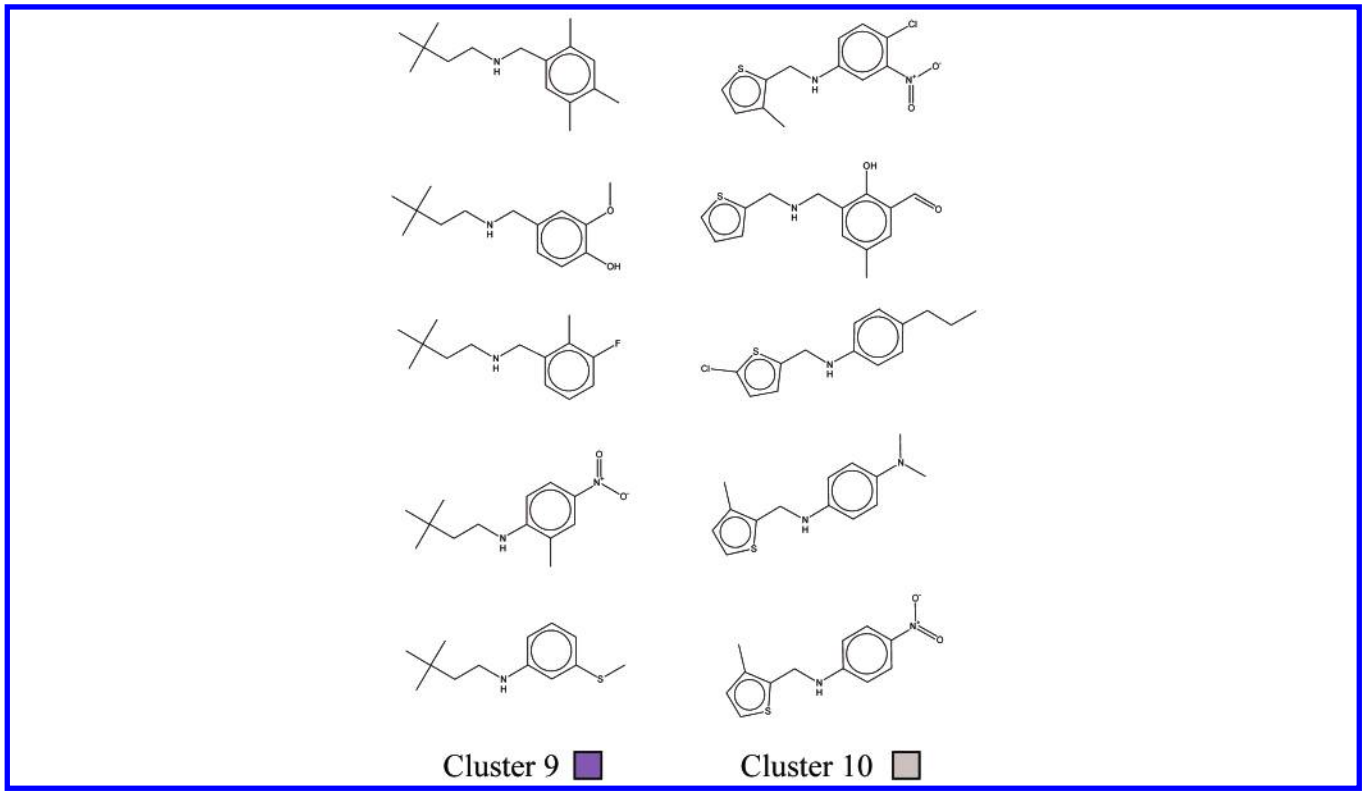


Figure 7. Chemical compounds comprising the 10 clusters highlighted in Figure 6.

that the atoms can move freely in Cartesian space. However, because of the rigidity of the covalent structure, which is implicitly captured by the rest of the data set, the atoms move in a concerted manner, restricted to a few modes of collective motion. The range of valid neighborhood radii is defined by two characteristic quantities: the local curvature of the manifold (upper bound) and the frequency at which it is sampled (lower bound). The latter decreases with increasing number of available conformations, asymptotically approaching zero at infinite sampling.

Many real-world data sets, however, do not have a clear manifold geometry. This may be due to noise or discontinuities in the observation space. Consider, for example, the combinatorial library in Figure 6. Here, the compounds are abstracted as points in a Euclidean space, such that the distances between the points represent the dissimilarity of the respective compounds in a particular molecular property space. This library is derived by combining a relatively small number of building blocks (100 amines and 100 aldehydes) in all possible combinations. While some of the resulting structures are closely related, there are clear chemical discontinuities resulting from the discrete nature of the descriptors and the diversity of the chemical fragments employed. These discontinuities, which represent distinct chemical classes, are not evident when we attempt to preserve all pairwise proximities (Figure 6a). On the contrary, there is significant loss of local information, causing scattering of closely related compounds and erroneous aggregation of unrelated ones. A similar effect was observed using a Sammon-like methodology that used a neighborhood radius determined from preexisting structure−activity data.[18] As the neighborhood radius decreases, structural families begin to emerge and become more clearly delineated, until we reach the fragmentation threshold. The "ideal" cutoff is one that represents a good compromise between the stress and the number of connected components (Figure 5). The resulting maps (Figure 6c) are visually compelling—they exhibit clusters that are consistent with the nature of the underlying descriptors, which in this case, encode predominantly the size, ring structure, branching, and heteroatom content of the molecules (Figure 7). As illustrated in Figure 5, the cutoff is not particularly sensitive to the embedding dimension, as one would expect from intuition.

IV. CONCLUSIONS

Owing to the highly organized structure of our physical world, chemical and biological data exhibit strong correlations, leading to patterns that lie on or close to a smooth low-dimensional manifold. The approach described in this work offers several important advantages: (1) it provides a reliable means for extracting the intrinsic dimensionality of a set of related patterns, (2) it produces informative visual representations that preserve the intrinsic clustering of the data, and (3) it offers a rigorous definition of chemical neighborhood by modeling the nonlinear geometry of the data space. Unlike previous approaches based on hit rates of high-throughput screening experiments[19] and structure−activity profiles,[20] our method defines similarity along the constrained surface of the underlying manifold. SPE can model nonlinear structures that are invisible to PCA and MDS, is simple to implement, and works effectively on a massive scale. Because of these characteristics, we expect this technique to be widely utilized in the chemical and biological sciences.

REFERENCES AND NOTES

(1) Johnson, M. A. G. M. M. *Concepts and applications of molecular similarity*; Wiley: New York, 1990.

(2) Bellman, R. E. *Adaptive Control Processes*; Princeton University: Princeton, 1961.
(3) Agrafiotis, D. K.; Lobanov, V. S.; Salemme, F. R. Combinatorial informatics in the post-genomics era. *Nature Rev. Drug Discovery* **2002**, *1*, 337−346.
(4) Livingstone, D. *Data analysis for chemists: applications to QSAR and chemical product design*; Oxford University Press: Oxford, 1996.
(5) Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24*, 417−441, 498−520.
(6) Borg, I.; Groenen, P. J. F. *Modern Multidimensional Scaling: Theory and Applications*; Springer: New York, 1997.
(7) Shepard, R. N.; Carroll, J. D. Parametric representation of nonlinear data structures. *International Symposium on Multivariate Analysis*; Academic Press: New York, 1965; pp 561−592.
(8) Martinetz, T.; Schulten, K. Topology representing networks. *Neural Networks* **1994**, *7*, 507−522.
(9) Sammon, J. W. A nonlinear mapping for data structure analysis. *IEEE Trans. Computers* **1969**, *18*, 401−409.
(10) Tenenbaum, J., B.; de Silva, V.; Langford, J., C. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290*, 2319−2323.
(11) Roweis, S., T.; Saul, L., K. Nonlinear dimensionality reduction by localy linear embedding. *Science* **2000**, *290*, 2323−2326.
(12) Agrafiotis, D. K.; Xu, H. A self-organizing principle for learning nonlinear manifolds. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 15869−15872.
(13) Cormen, T. H.; Leiserson, C. E.; Rivest, R. L. *Introduction to Algorithms*; The MIT Press McGraw-Hill Book Company: Cambridge, 1990.
(14) Xu, H.; Izrailev, S.; Agrafiotis, D. K. Conformational sampling by self-organization. submitted.
(15) Crippen, G. M.; Havel, T. F. *Distance geometry and molecular conformation*; Research Studies Press: Somerset, UK, 1988.
(16) Hall, L. H.; Kier, L. B. The molecular connectivity chi indexes and kappa shape indexes in structure−property modeling. *Reviews in Computational Chemistry*; VCH Publishers: New York, 1991; pp 367−422.
(17) Agrafiotis, D. K.; Bone, R. F.; Salemme, F. R.; Soll, R. M. System and method for automatically generating chemical compounds with desired properties; United States Patent 5,463,564, 1995.
(18) Clark, R. D.; Patterson, D. E.; Soltanshahi, F.; Blake, J. F.; Matthew, J. B. Visualizing substructural fingerprints. *J. Mol. Graphics Modelling* **2000**, *18*, 404−411.
(19) Brown, R. D.; Martin, Y. C. Use of structure−activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(20) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of molecular diversity descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

CI025631M