An Efficient Projection Protocol for Chemical Databases: Singular Value Decomposition Combined with Truncated-Newton Minimization

Dexuan Xie,[†,‡] Alexander Tropsha,[§] and Tamar Schlick*[,†]

Departments of Chemistry and Mathematics, Courant Institute of Mathematical Sciences, New York University and the Howard Hughes Medical Institute, 251 Mercer Street, New York, New York 10012

Received June 28, 1999

A rapid algorithm for visualizing large chemical databases in a low-dimensional space (2D or 3D) is presented as a first step in database analysis and design applications. The projection mapping of the compound database (described as vectors in the high-dimensional space of chemical descriptors) is based on the singular value decomposition (SVD) combined with a minimization procedure implemented with the efficient truncated-Newton program package (TNPACK). Numerical experiments on four chemical datasets with real-valued descriptors (ranging from 58 to 27 255 compounds) show that the SVD/TNPACK projection duo achieves a reasonable accuracy in 2D, varying from 30% to about 100% of pairwise distance segments that lie within 10% of the original distances. The lowest percentages, corresponding to scaled datasets, can be made close to 100% with projections onto a 10-dimensional space. We also show that the SVD/TNPACK duo is efficient for minimizing the distance error objective function (especially for scaled datasets), and that TNPACK is much more efficient than a current popular approach of steepest descent minimization in this application context. Applications of our projection technique to similarity and diversity sampling in drug design can be envisioned.

1. INTRODUCTION

The dramatic growth of chemical databases in recent years—largely due to advances in combinatorial chemistry and high throughput screening[5]—demands new methods for database analysis and representation. Compounds in chemical databases are conventionally characterized by "molecular descriptors" that reflect chemical connectivity, charge distribution, shape, physical attributes, and other properties. From a geometrical viewpoint, each compound is described as an *m*-dimensional vector whose coordinators are molecular descriptors. Therefore, analyses of chemical databases involve clustering, similarity, or dissimilarity sampling of multidimensional vector objects.

In many applications of database analysis, it is important to study the distance relationships among the compounds (points) in the dataset. Such analyses can be facilitated by mapping these compounds from the high-dimensional space onto a two- or three-dimensional (2D or 3D) vector space so that the clustering patterns (distance relationships) can be observed visually. The projection mapping is often formulated as a distance−geometry problem: find *n* points in 2D (or 3D) so that their interpoint distances match the corresponding values from the *m*-dimensional space as closely as possible. Since this problem is typically over-determined—there are $n(n-1)/2$ distances but only $2n$ ($3n$) Cartesian coordinates for a system of *n* compounds—an optimal approximate projection mapping is sought. This involves defining and minimizing a distance error objective function.

The distance−geometry problem has many important applications in molecular structure studies,[8,15,16] but it is difficult to solve. Only a local solution can usually be obtained, and a good starting point is important. See refs 6, 9, and 11 for theoretical analysis and numerical algorithms related to the distance−geometry problem.

Recently, a distance−geometry approach has been applied to the analysis and 2D projection mapping of molecular databases.[4,12,17] The algorithm was classified as the nonlinear mapping[17] or Sammon[4] method. The Sammon method uses the steepest descent (SD) minimization algorithm and a randomly chosen starting point. This approach may suffer from slow convergence and may generate a 2D mapping that poorly approximates the original distances when the number of compounds is large. Thus, finding a good initial guess for the mapping is an important and difficult objective.

In this paper, we define a low-dimensional projection mapping by the *singular value decomposition* (SVD),[10] a technique used for data compression in many practical applications such image processing and code deciphering. It is a factorization of rectangular matrices that reduces to the usual spectral (eigenvalue) decomposition when the matrices are square. This factorization, in contrast to optimization, only requires the input (high-dimensional) data vectors; no initial projection guess is needed. We assess the accuracy of the projection according to its level of approximation to the original intercompound distance values

* To whom correspondence should be addressed. Phone: (212) 998-3116. Fax: (212) 995-4152. E-mail: schlick@nyu.edu.

[†] New York University and the Howard Hughes Medical Institute.

[‡] Present address: Department of Mathematics and Ph.D. Program in Scientific Computing, University of Southern Mississippi, Hattiesburg, MS 39406-5045.

[§] School of Pharmacy, University of North Carolina, Chapel Hill, NC 27599-7360.

10.1021/ci990333j CCC: $19.00 © 2000 American Chemical Society
Published on Web 12/01/1999

in the *m*-dimensional space. We find that the accuracy of the SVD mapping depends on the distribution of the singular value magnitudes: if the first two singular values are much larger than the others, the 2D mapping has a high accuracy. This generalizes to mapping in higher dimensions as well; that is, if the first 10 singular values can be largely separated from the rest, a 10D projection can be accurate.

To achieve higher accuracy than obtained by the projection alone, we supplement the SVD projection by minimization of a new distance error objective function. This minimization is performed with TNPACK (our truncated-Newton program package).[18,21] With the SVD mapping as the starting point, TNPACK can efficiently generate a low-dimensional mapping that approximates the original distance relationships. The minimization component is especially important when SVD alone does not provide a sufficiently accurate mapping (in terms of matching the original distance relationships).

Numerical experiments are reported on four chemical datasets (ranging from 58 to 27 255 compounds). These datasets represent compounds with different types of biological activity. They include estrogens (ESTR), an artificial dataset made of eight groups of compounds with different phramacological activities (ARTF), monoaminooxidase inhibitors (MAO), and pesticides (PEST). All compounds in these datasets have been characterized with topological descriptors. In addition, the MAO dataset has also been characterized by binary descriptors (MAO_{01}). Each compound vector in these datasets has 312 components except for ESTR, which has $m = 308$. For these datasets, we report results for both scaled and unscaled data. While scaling remains an unresolved issue in the field,[23] we scaled the descriptors to the same unit range scale assuming that no one descriptor dominates the overall distance measures.

We find that, for the unscaled data, the SVD alone generates an excellent 2D projection: For ESTR, ARTF, and MAO, about 99% of the distance segments are within 10% of the original distances; for PEST (27 255 compounds), the corresponding value is 74%. Moreover, SVD is very fast: the computational time ranges from 0.08 to 8 min for these four datasets (on an SGI R10000 processor). The decomposition has a complexity of order $O(n^2m)$ floating point operations and $O(nm)$ memory locations. Though TNPACK offers only marginal improvements when the SVD mapping alone has a high accuracy, it is very fast, requiring from 1 s to 27 min for the ESTR, ARTF, and MAO datasets, respectively. TNPACK minimization is also far more efficient than the steepest descent minimization (38 times faster for ESTR and about 110 times faster for MAO), the method used in refs 4 and 17.

For the scaled datasets, of which the binary MAO_{01} database serves as an extreme case, we find in contrast that it is much more difficult to define a satisfactory 2D mapping by SVD alone. Since the first two SVD values are usually not much larger than the others in the scaled cases, the resulting 2D SVD projections can be poor approximations. However, the TNPACK minimizations that follow SVD become crucial. For example, the 2D SVD mapping of MAO_{01} only has about 0.004% of the distance segments within 10% of the original distance values. TNPACK increased this number to 30% in 1 min of CPU time. For the other scaled datasets, the 2D SVD/TNPACK mappings have up to about 43% and 80% of the distance segments within 10% and 20% of the original distance values, respectively. We also show that the accuracy of SVD and SVD/TNPACK projections can be improved sharply when the dimension number of the projection space is increased from two to five or ten.

Finally, analysis of the 2D projections in terms of chemical structure similarity reveals that points that are close together in the 2D mapping have similar chemical structures and vice versa.

The remainder of the paper is organized as follows. In section 2, we formulate the mapping problem and the projection assessment in terms of distance functions (errors). Sections 3 and 4 outline, respectively, the SVD-based mapping and our distance−geometry algorithm using TNPACK to supplement the SVD mapping. Section 5 presents the numerical results and preliminary chemical structure analyses for both scaled and unscaled datasets. Conclusions are summarized in section 6.

2. FORMULATION OF THE MAPPING PROBLEM

We express the dataset S as a collection of n vectors

$$S = \{X_1, X_2, ..., X_n\}$$

where each vector $X_i = (x_{i1}, x_{i2}, ..., x_{im})^T$ consists of the m descriptors $\{x_{ik}\}$, which are real numbers. We now define the distance quantities δ_{ij} and d_{ij} corresponding to vectors in the original (R^m) and projected (e.g., R^2) spaces.

From those descriptors, similarity between each pair of compounds X_i and X_j can be described by the following *Euclidean distance*:

$$\delta_{ij} = [\sum_{k=1}^{m}(x_{ik} - x_{jk})^2]^{1/2} \quad (1)$$

There are $n(n-1)/2$ distance segments $\{\delta_{ij}\}$ in S for pairs $i < j$.

Assume we have a mapping from R^m to R^{low} that takes each point $X_i \in R^m$ to $Y_i \in R^{low}$, where $low \ll m$. Typically the integer low is 2 or 3, but we use $low = 10$ in some cases discussed below; the projection cannot be easily visualized for $low > 3$, but the compressed matrix $\tilde{S}$ from S ($n \times low$ instead of $n \times m$) can be useful for other database applications. The corresponding interpoint distances for the vectors $Y_i = (y_{i1}, y_{i2}, ..., y_{ilow})^T$ is denoted as $d(Y_i,Y_j)$, where

$$d(Y_i,Y_j) = [\sum_{k=1}^{low}(y_{ik} - y_{jk})^2]^{1/2}$$

An ideal mapping will generate points $\{Y_i\}$ that match the original values, i.e., satisfy

$$d(Y_i,Y_j) = \delta_{ij} \quad (2)$$

for all pairs $i < j$. However, no such mapping exists in general because the problem is typically overdetermined—finding $n \times low$ unknowns $\{y_{ik}\}$ satisfying $n(n-1)/2$ equations of form (2). An optimal approximate mapping is thus sought.[16]

We now define various error measures to assess the relationship between the original and projected pairwise distances. Following ref 16, we use the relative error

expression as one measure of the quality of the approximation of $d(Y_i,Y_j)$ to δ_{ij}:

$$|d(Y_i,Y_j) - \delta_{ij}| \le \epsilon\delta_{ij} \quad \text{when} \quad \delta_{ij} > d_{\min} \tag{3}$$

$$d(Y_i,Y_j) \le \tilde{\epsilon} \quad \text{when} \quad \delta_{ij} \le d_{\min} \tag{4}$$

where ϵ, $\tilde{\epsilon}$, and $d_{\min}$ are given small positive numbers of less than 1. For example, we set $\epsilon = 0.1$ to specify a 10% accuracy ($d_{\min} = 10^{-12}$ and $\tilde{\epsilon} = 10^{-8}$). The second case above (very small original distance) may occur when two compounds in the datasets are highly similar.

The total number T_d of the distance segments $d(Y_i,Y_j)$ satisfying (3) or (4) reflects the distance preservation of our mapping. We define the percentage ρ of the distance segments satisfying (3) or (4) as

$$\rho = \frac{T_d}{n(n-1)/2} \times 100 \tag{5}$$

The greater the ρ values, the better the mapping and the more information can be inferred from the projected views of the complex data (assuming the original distances are meaningful).

We also use the following "average relative error" to assess the mapping:

$$\text{error}_r = \frac{\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{|d(Y_i,Y_j) - \delta_{ij}|^2}{\delta_{ij}^{\,2}}\right)^{1/2}}{n(n-1)/2} \tag{6}$$

where δ_{ij} in the denominator is replaced by 1 if $\delta_{ij} \le d_{\min}$. In addition, we consider the "average absolute error" used in ref 17:

$$\text{error}_a = \frac{\left(\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} |d(Y_i,Y_j) - \delta_{ij}|^2\right)^{1/2}}{n(n-1)/2} \tag{7}$$

The objective function E for distance refinement used in our minimization phase is defined in section 4 (see 15)).

3. SVD-BASED PROJECTION

SVD is a technique for data compression used in image processing and code deciphering. Essentially, SVD is a fractorization for rectangular matrices that reduces to the eigenvalue decomposition when the matrices are square. SVD defines two orthogonal coordinate systems (for the domain and range of the matrix) with corresponding singular values $\{\sigma_i\}$. Data compression can be achieved by expressing the matrix elements in terms of the components corresponding to κ nonzero singular values rather than all r values (r is the matrix rank). When those κ singular values ($\sigma_1 \ge \sigma_2 \ge ... \ge \sigma_\kappa$) are significantly larger than the remaining ones ($\sigma_{\kappa+1} \ge \sigma_{\kappa+2} \ge ... \ge \sigma_r$), the rank-$\kappa$ approximation is good.

In our context, we construct a rectangular matrix X by listing, in rows, the m descriptors of the n dataset compounds:

$$X = (X_1, X_2, ..., X_n)^T = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1m} \\ x_{21} & x_{22} & ... & x_{2m} \\ \vdots & \vdots & ... & \vdots \\ x_{n1} & x_{n2} & ... & x_{nm} \end{bmatrix}$$

Typically, $n \gg m$ for large datasets. The SVD decomposition of X can be written as

$$X = U\Sigma V^T$$

where $U_{n\times n}$ and $V_{m\times m}$ are the orthogonal matrices:

$$U = (u_1, u_2, ..., u_n) \quad \text{and} \quad V = (v_1, v_2, ..., v_m)$$

where $u_i \in R^n$ and $v_i \in R^m$. The diagonal matrix $\Sigma_{n\times m}$ contains the singular values arranged in decreasing order:

$$\Sigma = \text{diag}\{\sigma_1, \sigma_2, ..., \sigma_m\}$$

where $\sigma_1 \ge \sigma_2 \ge ... \ge \sigma_r$ and $\sigma_{r+1} = ... = \sigma_m = 0$. Thus, X can be written as the sum of rank-1 matrices:

$$X = \sum_{k=1}^{r} \sigma_k u_k v_k^T \tag{8}$$

A rank-κ approximation can be written as[10]

$$X^\kappa = \sum_{k=1}^{\kappa} \sigma_k u_k v_k^T \qquad ||X^\kappa - X|| = \sigma_{\kappa+1}$$

Let $u_k = (u_{1k}, u_{2k}, ..., u_{nk})^T$ and e_i be an $n \times 1$ unit vector with 1 in the ith component and 0 elsewhere. The inner product $u_k^T e_i = u_{ik}$. Using (8), we then express each vector X_i as a linear combination of vectors $\{v_k\}_{k=1}^m$:

$$X_i = X^T e_i = \sum_{k=1}^{r} \sigma_k (u_k^T e_i) v_k = \sum_{k=1}^{m} \sigma_k u_{ik} v_k, \quad i = 1, 2, ..., n$$

where $\sigma_{r+1} = ... = \sigma_m = 0$. Hence, with $\{v_k\}_{k=1}^m$ as a new orthonormal basis of R^m, we express X_i in terms of new coordinates

$$X_i = (\sigma_1 u_{i1}, \sigma_2 u_{i2}, ..., \sigma_r u_{ir}, 0, ..., 0)^T$$

The corresponding Euclidean norm $||\cdot||$ of X_i, $||X_i||^2 = \sum_{k=1}^m x_{ik}^2$, is written as

$$||X_i||^2 = \sum_{k=1}^{r} (\sigma_k u_{ik})^2 \tag{9}$$

We now use SVD to define the *low*-dimensional mapping vector Y_i of X_i as the natural projection of X_i onto the subspace spanned by the basis vectors $v_1, v_2, ..., v_{low}$

$$Y_i = \sum_{k=1}^{low} \sigma_k u_{ik} v_k$$

for each compound i, so that Y_i can also be written as a vector of R^{low}:

$$Y_i = (\sigma_1 u_{i1}, \sigma_2 u_{i2}, ..., \sigma_{low} u_{ilow}) \tag{10}$$

The projected vector Y_i is a *low*-dimensional approximation of X_i with the following relative error expression:

$$\zeta_i = \frac{||X_i - Y_i||}{||X_i||} = \left[1 - \left(1 + \frac{\sum_{k=low+1}^{r} (\sigma_k u_{ik})^2}{\sum_{k=1}^{low} (\sigma_k u_{ik})^2}\right)\right]^{1/2} \quad (11)$$

When the first *low* singular values dominate the rest, the term $\sum_{k=1}^{low}(\sigma_k u_{ik})^2$ becomes much larger than the term $\sum_{k=low+1}^{r}(\sigma_k u_{ik})^2$, resulting in small relative errors ζ_i for all $1 \le i \le n$.

Let $\bar{S}$ be the collection of these n mapped vectors from S: $\bar{S} = \{Y_1, Y_2, ..., Y_n\}$. The error approximation of $\bar{S}$ with respect to the original dataset S can be described by the average relative error ζ:

$$\zeta = (\sum_{i=1}^{n} \zeta_i)/n \quad (12)$$

If the above average relative error ζ is sufficiently small (such as $\zeta \le 0.1$), the mapping $\bar{S}$ defined by SVD preserves approximately the distance relationships among the n chemical compounds $\{X_k\}$ of S; otherwise, this projection can serve as a starting point for further refinement. Even this achievement is significant since selecting the starting point is difficult in distance−geometry problems. It is often selected randomly,[16,17] leading to large computational time for convergence and affecting the local solution obtained.

Further, we can write the distance segments $d_{ij} = ||Y_i - Y_j||$ and $\delta_{ij} = ||X_i - X_j||$ as follows:

$$d_{ij} = [\sum_{k=1}^{low} \sigma_k{}^2 (u_{ik} - u_{jk})^2]^{1/2} \quad \text{and}$$

$$\delta_{ij} = [\sum_{k=1}^{r} \sigma_k{}^2 (u_{ik} - u_{jk})^2]^{1/2}$$

Clearly, d_{ij} is an approximation of δ_{ij} with the following relative error expression:

$$\frac{\delta_{ij} - d_{ij}}{\delta_{ij}} = 1 - \left[1 + \frac{\sum_{k=low+1}^{r} \sigma_k{}^2 (u_{ik} - u_{jk})^2}{\sum_{k=1}^{low} \sigma_k{}^2 (u_{ik} - u_{jk})^2}\right]^{-1/2} \quad (13)$$

where $d_{ij} \le \delta_{ij}$, and δ_{ij} is assumed to be positive. Thus, if the first *low* singular values dominate the rest, the above relative error becomes small for all $i < j$, implying that the *low*-dimensional distance d_{ij} is a good approximation of the original distance δ_{ij}.

4. TNPACK REFINEMENT FOLLOWING THE SVD PROJECTION

When the percentage ρ defined in (5) (or the error ζ defined in (12)) is not satisfactory, we formulate an objective function E and minimize it by our truncated-Newton program package, TNPACK.[18,21] Specifically, we seek a projection refinement taking Y^* to Y such that

$$E(Y^*) = \min_{Y \in R^{low \cdot n}} E(Y) \quad (14)$$

where the objective function $E(Y)$ is defined by

$$E(Y) = (1/4)\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \omega_{ij} |d(Y_i, Y_j)^2 - \delta_{ij}{}^2|^2 \quad (15)$$

Here, $Y = (Y_1, Y_2, ..., Y_n)^T \in R^{low \cdot n}$ with $Y_i = (y_{i1}, y_{i2}, ..., y_{ilow})$ for $i = 1, 2, ..., n$, and the weights $\{\omega_{ij}\}$ are set as $\omega_{ij} = 1/\delta_{ij}{}^4$ if $\delta_{ij}{}^4 \ge \eta$ and $\omega_{ij} = 1$ if $\delta_{ij}{}^4 < \eta$. The parameter η is a small positive number such as 10^{-12}. Equations similar to (15) have been reported (for example, in ref 17) with various weight choices.

Since E is a simple polynomial function of Y, its derivatives are well defined at every vector of $R^{low \cdot n}$, and a second-derivative method such as TNPACK can be applied efficiently. One of the features of TNPACK is an application-tailored preconditioner matrix (that approximates the Hessian of the objective function) used to accelerate convergence.[20] For the present applications, however, we only used a simple diagonal preconditioner, namely, terms $\partial^2 E(Y)/\partial y_{ik}{}^2$ ($i = 1, 2, ..., n$ for $k = 1, 2, ..., low$).

Various objective functions have been reported in the literature for the distance−geometry problem. A typical one follows:

$$E(Y) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \omega_{ij} |d(Y_i, Y_j) - \delta_{ij}|^2 \quad (16)$$

where $\{\omega_{ij}\}$ are weights (refs 4 and 16, for example). Using our E of (15) instead of E defined in (16) can be numerically advantageous since the potential problem of a near-zero denominator ($d(Y_i, Y_j)$ term in $\partial E(Y)/\partial y_{ik}$) can be avoided. The two functions are also closely related. For $\delta_{ij} > 0$, we have

$$\frac{|d(Y_i, Y_j) - \delta_{ij}|^2}{\delta_{ij}{}^2} = \frac{|d(Y_i, Y_j) - \delta_{ij}|^2 |d(Y_i, Y_j) + \delta_{ij}|^2}{\delta_{ij}{}^2 |d(Y_i, Y_j) + \delta_{ij}|^2}$$

$$= \frac{|d(Y_i, Y_j)^2 - \delta_{ij}^2|^2}{\delta_{ij}{}^2 |d(Y_i, Y_j) + \delta_{ij}|^2} \le \frac{|d(Y_i, Y_j)^2 - \delta_{ij}{}^2|^2}{\delta_{ij}{}^4}$$

Hence, if

$$\frac{|d(Y_i, Y_j)^2 - \delta_{ij}{}^2|^2}{\delta_{ij}{}^4} \le \epsilon^2$$

the relative error inequality (3) holds.

5. NUMERICAL EXAMPLES

The compound datasets used for testing our mapping approach are as follows: ESTR ($n = 58$ and $m = 308$), MAO ($n = 1623$ and $m = 312$), MAO_{01} ($n = 1623$ and $m = 153$, binary descriptors), and PEST ($n = 27\,255$ and $m = 318$). We also constructed dataset ARTF ($n = 402$ and $m = 312$) by merging eight different groups of molecules with different types of pharmacological activities. Binary descriptors for MAO_{01} were generated from the software MACCS II.[14] Descriptors for PEST and other datasets were calculated by

Table 1. Performance of the 2D SVD and SVD/TNPACK (TN) Mappings[a]

dataset	E		$\|\|g\|\|$		no. of TN iterations	CPU time	
	SVD	TN	SVD	TN		SVD	TN
			Unscaled				
ESTR	1.29	0.024	2.65×10^{-3}	9.12×10^{-6}	50	0.08 s	1.03 s
ARTF	45.8	3.04	4.98×10^{-3}	3.21×10^{-5}	72	2.66 s	1.95 min
MAO	573.3	298.3	2.75×10^{-2}	2.81×10^{-3}	83	7.62 s	24.08 min
PEST						8.41 min	
			Scaled				
ESTR	2.18×10^{2}	6.53×10^{1}	2.17×10^{2}	3.86×10^{-4}	20	0.07 s	0.40 s
ARTF	8.66×10^{4}	2.92×10^{3}	1.49×10^{3}	1.92×10^{-2}	57	1.16 s	25.31 s
MAO	1.82×10^{5}	6.06×10^{4}	8.35×10^{3}	4.54×10^{-1}	201	5.62 s	24.66 min
MAO_{01}	2.4×10^{5}	9.79×10^{4}	3.79×10^{3}	9.59×10^{-1}	11	3.65 s	46.46 s
PEST						6.76 min	

[a] E is the minimization objective function defined in (15).

using the software package Molconn-Z[1] and Molconn-X,[2] respectively.

We considered both unscaled and scaled datasets for our analyses. We scaled the descriptors $\{x_{ij}\}$ to the same unit range scale using the following formula for each column j:

$$\hat{x}_{ij} = \frac{x_{ij} - x_{\min,j}}{x_{\max,j} - x_{\min,j}}, \quad 1 \le i \le n \tag{17}$$

where $x_{\min,j} = \min_{1 \le i \le n} x_{ij}$ and $x_{\max,j} = \max_{1 \le i \le n} x_{ij}$. The scaling procedure (17) is often referred to as a standardization of descriptors. It assumes that no one descriptor dominates the overall distance measures, and is widely used in practice. We have also considered a different scaling procedure in ref 22, where we found results to be slightly better than those reported here. Throughout the figures we use the superscript S notation with the dataset name to indicate scaling; no superscript implies raw data (unscaled).

For the SVD procedure, we used the NAG library.[3] For simplicity, we used all default parameters of TNPACK[18,21] for the minimization that follows the projection. The target accuracy ϵ in (3) was set to 0.1 and 0.2 (for some tests of scaled datasets). The termination rule in TNPACK for the iterates $\{Y^k\}$ is defined as

$$||g(Y^k)|| < \epsilon_g(1 + E(Y^k)) \tag{18}$$

where $\epsilon_g = 10^{-5}$, and g is the gradient vector of E. All computations were performed in double precision on a single R10000 195 MHz processor of an SGI Power Challenge L computer at New York University.

5.1. Performance of the 2D SVD and SVD/TNPACK Mappings. Table 1 displays the performance of the 2D SVD and SVD/TNPACK mappings for both scaled and unscaled datasets (including the binary dataset). Because of memory limitations, TNPACK was not applied to the large dataset PEST. From Table 1 we see that both the SVD and SVD/TNPACK duo are efficient in generating the 2D mappings. The longest SVD CPU time is about 8 min for PEST (n = 27 255), and TNPACK only took 2 min for the ARTF dataset (n = 402) to calculate a minimum point of E.

5.2. Accuracy of the 2D SVD and SVD/TNPACK Mappings. Table 2 lists five different error assessments to the 2D SVD and SVD/TNPACK mappings. Here ζ is defined in (12), indicating an average relative error of each 2D SVD mapping point with respect to the corresponding chemical compound. The percentage ρ is defined by (5) to indicate the portion of the distance segments that satisfy error requirement (3), that is, are within 10% of the original distance values. error_a is the average absolute error defined in (7), used in ref 17. error_r is the average relative error defined in (6). error_{rel} is the standard relative error defined as follows:

Table 2. Error Estimates of the 2D SVD and SVD/TNPACK Mappings[a]

dataset	ζ	ρ	error_{rel}	error_r	error_a
			SVD Alone for Unscaled Datasets		
ESTR	0.08	98.97	6.04×10^{-4}	1.25×10^{-3}	2.57×10^{-1}
ARTF	0.05	98.97	2.69×10^{-4}	1.55×10^{-4}	4.19×10^{-2}
MAO	0.12	99.48	3.92×10^{-4}	2.75×10^{-5}	6.92×10^{-3}
PEST	0.087	74.00	0.1620		
			SVD Alone for Scaled Datasets		
ESTR	0.38	8.47	0.4202	1.37×10^{-2}	4.22×10^{-2}
ARTF	0.34	25.38	0.3090	1.72×10^{-3}	4.63×10^{-3}
MAO	0.34	1.81	0.5056	4.98×10^{-4}	1.19×10^{-3}
MAO_{01}	0.70	0.004	0.6477	5.77×10^{-4}	3.47×10^{-3}
PEST	0.15	11.79	0.6314		
			SVD/TNPACK Duo for Scaled Datasets		
ESTR		38.54	0.2047	6.29×10^{-3}	2.05×10^{-2}
ARTF		43.26	0.1679	8.73×10^{-4}	2.52×10^{-3}
MAO		38.72	0.2397	2.25×10^{-4}	5.65×10^{-4}
MAO_{01}		29.10	0.3729	3.46×10^{-4}	1.99×10^{-3}

[a] Here ζ, ρ, error_{rel}, error_r, and error_a are defined in (12), (5), (19), (6), and (7), respectively.

$$\text{error}_{rel} = \left[\left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |d(Y_i,Y_j) - \delta_{ij}|^2 \right) / \left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \delta_{ij}^2 \right) \right]^{1/2} \tag{19}$$

The smaller these error measurements, the more useful the 2D projection is for subsequent database analysis.

We note from Table 2 that all error values of the SVD mapping are small for the four unscaled datasets. Three of them have percent values ρ of about 99 that describe the portion of distance segments satisfying the error requirement (3). Hence, the SVD mapping can be considered satisfactory in preserving the distance relationships in the high-dimensional space.

Scaling, especially the binary dataset, however, makes the distance preservation inherently difficult. From Table 2 we see that the 2D mappings defined by SVD alone have poor accuracies for the scaled cases. TNPACK greatly improved

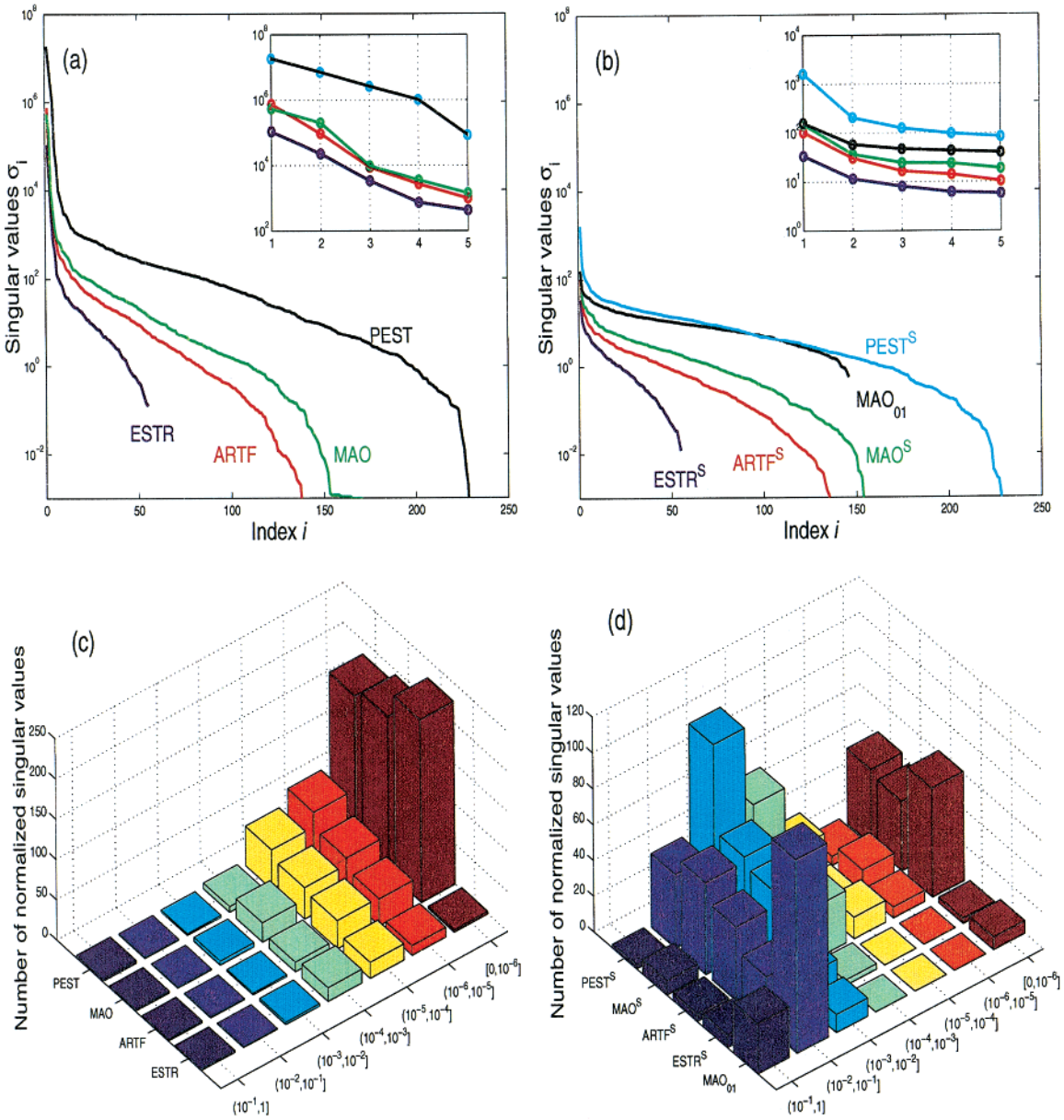

Figure 1. Singular values and their distributions for ESTR, ARTF, MAO, PEST, and MAO_{01}, unscaled (left) and scaled (right, marked by superscripts S).

the projection in this regard (ρ ranges from 30% to 43%), though it is still less accurate compared to the ones in the unscaled case.

Figure 1 displays the singular values and their distributions for all four datasets. From Figure 1a we see that the first two singular values of the unscaled datasets ESTR, ARTF, MAO, and PEST are much larger than the others. In contrast, for the scaled datasets, including the binary dataset MAO_{01}, the first two singular values are not significantly larger than the others (Figure 1b).

To fairly compare singular values for the different datasets, we also normalized the singular values using the following formula: $\hat{\sigma}_i = \sigma_i/\max_{1\le j\le r}\sigma_j$. Parts c and d of Figure 1 compare the distributions of the normalized singular values σ_i on seven intervals: $(10^{-k}, 10^{-(k-1)}]$ for $k = 1-6$ and $[0, 10^{-6}]$ for the unscaled and scaled datasets, respectively. The same trend is evident. This explains why the 2D mapping is good for the unscaled datasets but poor for the scaled datasets.

Figure 2 shows that the accuracy of the SVD and SVD/TNPACK mappings for the scaled dataset can be improved sharply when the number of dimensions (*low*) of the projection space is increased from two to ten. In particular, SVD alone defines a satisfactory 2D mapping when the dimension of the projection space is sufficiently large. Here we analyzed two scaled datasets ($ESTR^S$ and $ARTF^S$) in terms of the percentage ρ defined in (5) to indicate the accuracy. The value of ρ is nearly doubled when $\eta = 0.2$ instead of 0.1 in (5). We also found it useful to use higher-order SVD mappings for the purpose of selecting initial points for minimization refinement. See ref 22 for an illustration of a 3D mapping for ARTF.

5.3. TNPACK vs SD. Table 3 compares the performance between TNPACK and the steepest descent method for minimizing $E(Y)$ for the datasets ESTR and MAO. SD has been used in similar applications of multidimensional data projection,[4,17] and hence this comparison is important in the present application context. Here the termination rule (18)

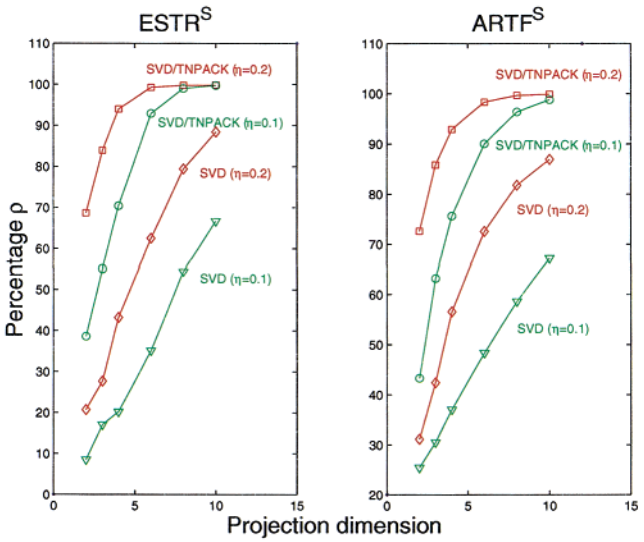

Figure 2. The percentage ρ defined in (5) increases with the number of dimensions of the projection space ($\eta = 0.1$ and 0.2 used).

Table 3. Comparison of TNPACK versus SD for Minimizing E

method	final E	final $\|\|g\|\|$	no. of iterations	CPU time
		Unscaled ESTR		
SD	0.024	1.02×10^{-5}	33 573	38.46 s
TNPACK	0.024	9.11×10^{-6}	50	1.03 s
		Unscaled MAO		
SD	276.14	2.77×10^{-3}	194 263	49.91 h
TNPACK	298.3	2.81×10^{-3}	83	0.40 h
		Scaled ESTR		
SD	6.70×10^{3}	6.79×10^{-4}	2 260	2.40 s
TNPACK	6.53×10^{1}	3.86×10^{-4}	20	0.40 s
		Scaled MAO		
SD	6.01×10^{4}	6.01×10^{-1}	7 856	2.04 h
TNPACK	6.06×10^{4}	4.54×10^{-1}	201	0.41 h

Table 4. Comparison of TNPACK Minimization Using the SVD Mapping as a Starting Point (SVD) versus a Randomly Selected Starting Point (RAN)

starting point X^0	final E	final $\|\|g\|\|$	no. of iterations	CPU time (min)
		Unscaled ARTF		
SVD	3.04	3.21×10^{-5}	72	1.95
RAN	1.42×10^{4}	1.23×10^{-2}	300	9.63
		Scaled ARTF		
SVD	2.92×10^{3}	1.92×10^{-2}	57	0.42
RAN	2.91×10^{3}	2.56×10^{-2}	112	1.33

and the SVD starting point were used by both TNPACK and SD. We see that TNPACK is much more efficient: 38 times faster than SD for the unscaled ESTR and about 110 times faster for the unscaled MAO to find a minimum point.

5.4. SVD Starting Point vs Random Starting Point. Table 4 also compares the performance of TNPACK using the SVD mapping as the starting point with that using a randomly selected starting point for dataset ARTF. The SVD starting point clearly helps accelerate the minimization process significantly.

5.5. Error Objective Function E vs $\mathcal{E}$. Figure 3 compares our objective function E with the typical objective function $\mathcal{E}$ given in (16). Here we set the weights $\omega_{ij} = 1/\delta_{ij}^2$ for $\delta_{ij} > 10^{-12}$ and $\omega_{ij} = 1$ for $\delta_{ij} \leq 10^{-12}$. The figure shows that

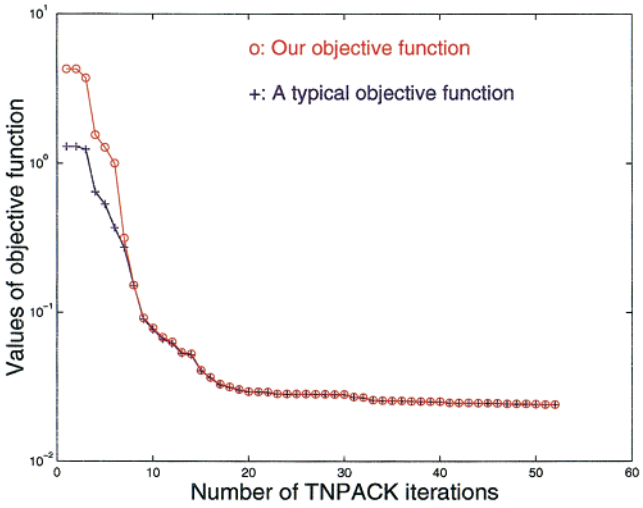

Figure 3. Comparison of our objective function E with the typical objective function $\mathcal{E}$ given in (16) for ESTR.

$\mathcal{E}(Y) \leq E(Y)$ for all Y, and $E(Y)$ becomes close to $\mathcal{E}(Y)$ as Y approaches a minimum point of $E(Y)$. Hence, the choice of the particular function for E between these two formulations is not important, and our choice (15) is computationally preferred.

5.6. 2D Mapping Displays. Figure 4 displays the 2D mappings for ESTR, ARTF, MAO, MAO_{01}, and PEST. These illustrations also compare the plots of the 2D SVD and SVD/TNPACK mappings for both scaled and unscaled datasets (except for PEST). In some cases, an inset shows a zoomed view. Since the SVD mappings for the unscaled ESTR, ARTF, and MAO already produce high accuracies (Table 2), TNPACK refinements change these projections only slightly. However, the SVD plots for the scaled datasets including the binary dataset MAO_{01} have been significantly changed by TNPACK to improve the distance values in 2D with respect to the original values.

5.7. Chemical Structure Similarity Analysis. Figure 5 displays the distribution of eight chemical/pharmacological classes of ARTF based on the 2D SVD/TNPACK mappings for the scaled and unscaled ARTF. The number of compounds in each class is indicated in the figure after the class name. We note that clusters corresponding to individual pharmacological subsets are generally very close to each other, though partial overlap of clusters is evident.

The ecdysteroids group forms a diverse but separate set of points. The estrogen class is also clustered and somewhat separate from the others. The strong overlap of the three clusters corresponding to D1 agonists, D1 antagonists, and H1 receptor ligands is reasonable given the relative chemical similarity of these compounds: all act at receptors of the same pharmacological class (i.e., G-protein coupled receptors). In fact, some of the H1 ligands have been initially tested for dopaminergic affinity and are also members of the other two groups. This explains the complete overlap of points in some cases.

For further chemical analysis, we select 14 compound representatives from the estrogen, DHFR, and AChE classes of ARTF, and display the corresponding submapping in Figure 6 for the unscaled and scaled data. Selected chemical structures are shown in Figure 7, plotted from 3D coordinates.

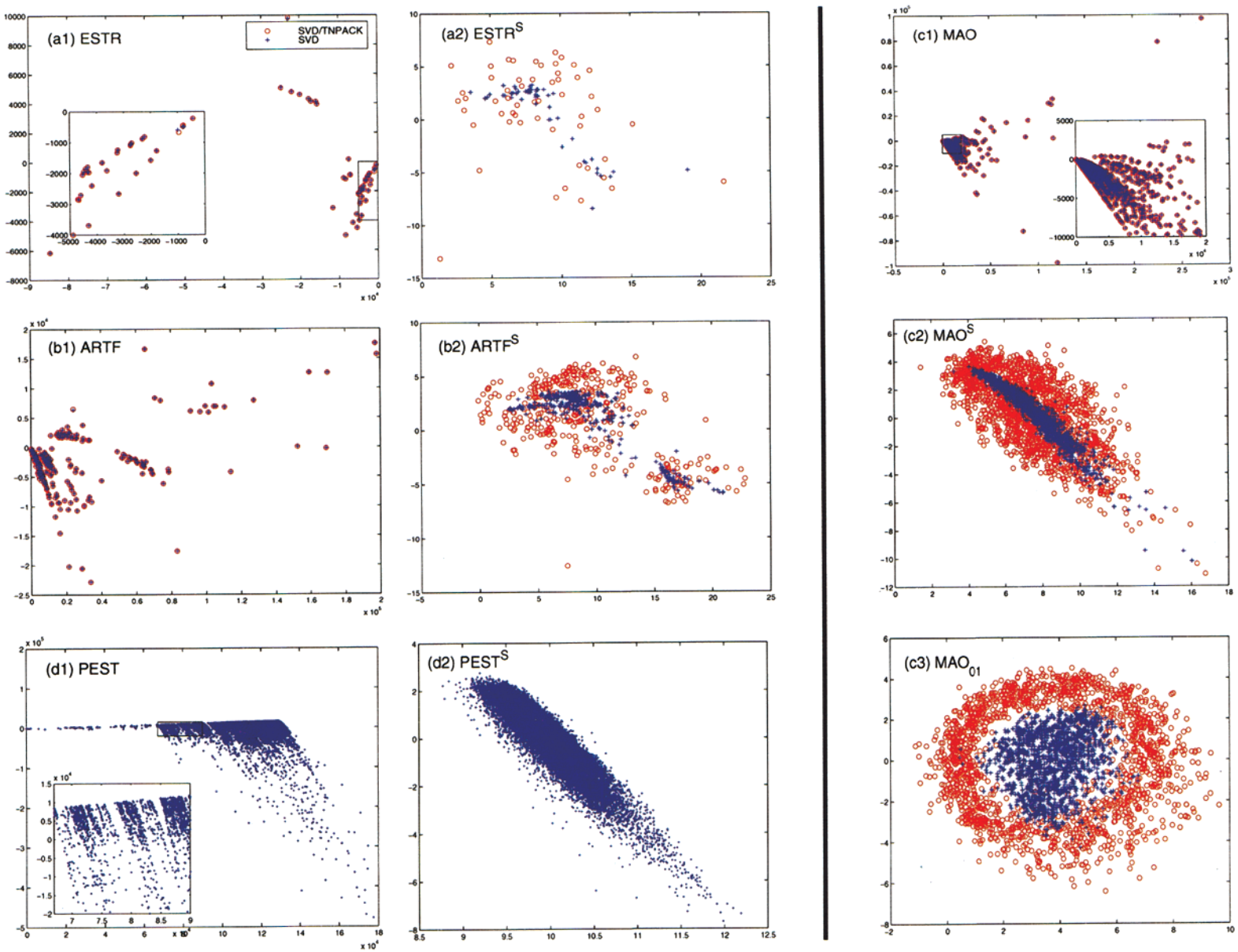

Figure 4. 2D mappings defined by SVD and SVD/TNPACK for ESTR ($n = 58$, $m = 308$), ARTF ($n = 402$, $m = 312$), MAO ($n = 1623$, $m = 312$), PEST ($n = 27\,255$, $m = 312$), and MAO_{01} ($n = 1623$, $m = 153$), unscaled and scaled (marked by superscripts S).

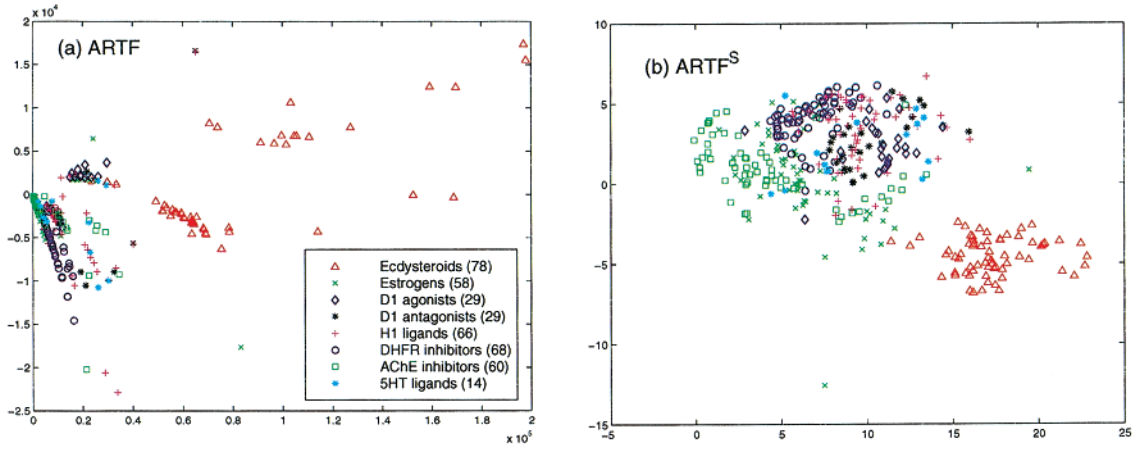

Figure 5. 2D SVD/TNPACK mappings of the eight pharmacological classes, unscaled and scaled (marked by superscript S).

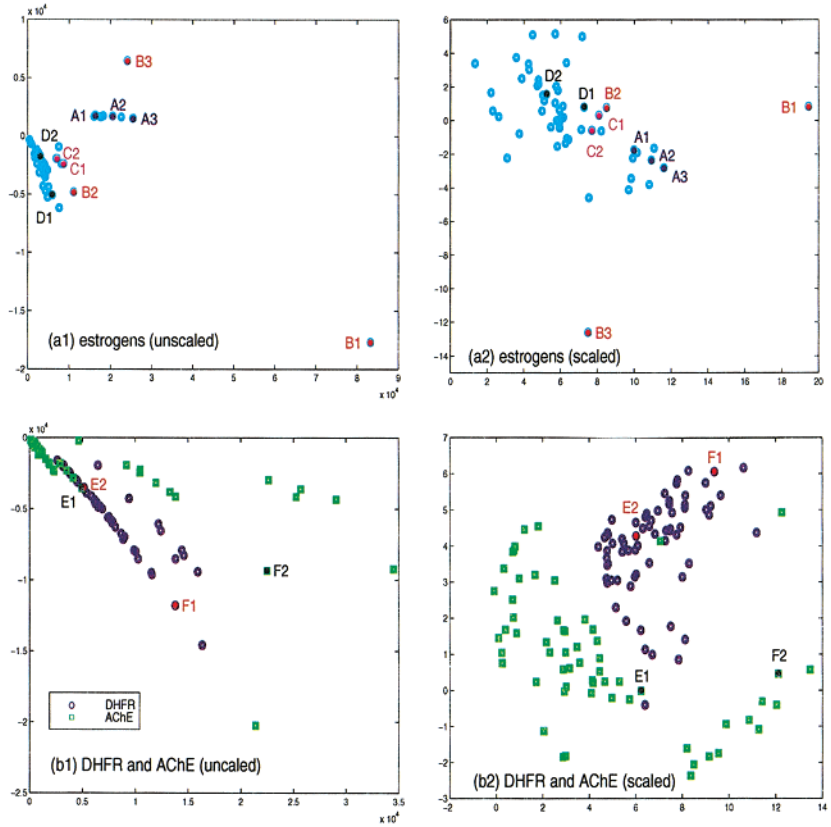

Figure 6. Chemical similarity and diversity of ARTF (see Figure 7 for chemical structures).

The 2D projections are characterized by several smaller subclusters and a few singletons. For the estrogens, we observe common structural motifs in the following subclusters: four-cyclic (points A1, A2, and A3), three-cyclic (C1, C2), and two-cyclic (D1, D2) molecules. Furthermore, the three singletons B1, B2, and B3 correspond to unique structures. This limited analysis, for both the scaled and unscaled projections, shows that compounds that belong to the same cluster are generally similar to each other yet less similar to compounds from other clusters. Further clustering analysis done in ref 22 also shows that points that are close in space in the 2D projection are similar, while those that are distant are dissimilar (diversity application).

Although the DHFR inhibitors and AChE inhibitors come from different pharmacological classes, we again observe (b1 or b2 in Figure 6) similar structures (e.g., E1 and E2), whereas distant points correspond to dissimilar structures (e.g., F1 and F2). The points E1 and E2, however, are close in the mapping corresponding to the unscaled data but far apart in the mapping corresponding to the scaled dataset. Here the relative errors of the SVD/TNPACK mapping between points E1 and E2 in both scaled and unscaled cases are 6.02×10^{-2} and 3.88×10^{-2}, respectively.

6. CONCLUSIONS

Our approach to low-dimensional mappings of chemical databases represented in the high-dimensional descriptor space is based on the SVD coupled with the TNPACK package. The former decomposes the compounds in terms of generalized eigenvectors whose associated singular values help analyze the degree of component coupling and dependence. When the first two singular values are large relative

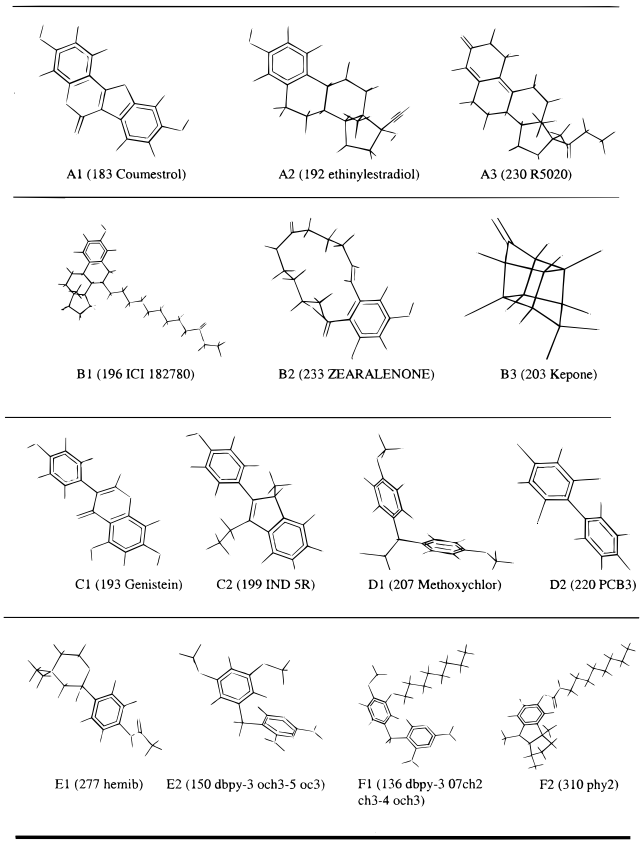

Figure 7. Chemical structures from database ARTF. The index and name are listed for each compound.

to the others, the 2D projections are fairly accurate in terms of matching the intercompound distances in the original space. The SVD projection is followed by minimization of an objective function to further reduce the amount of distance discrepancies. This quartic objective function measures the sum of discrepancies for each pair of compounds (i and j) between the 2D value (d_{ij}) and the m-dimensional value (δ_{ij}). TNPACK offers an efficient approach for minimization, and the minimization is rapidly convergent when a reasonable initial guess—from the SVD projection—is supplied. The amount of refinement of the SVD projection depends on the initial value of the objective function: in some cases, minimization leads to minor improvements and in others (e.g., scaled datasets) to more substantial function reduction. The SVD component is very fast, even for large datasets, and the truncated-Newton minimization is very efficient compared to the commonly-used steepest descent method in this context;[4,17] we have not implemented TNPACK for the dataset PEST (27 255 compounds) due to memory restrictions, but this limitation might be lifted in the future with a low-memory variant of our minimization algorithm. Comparison of our distance−geometry approach to a neural network procedure by Kohonen[13] in our related work[22] is also favorable.

In the case of the scaled (including binary) descriptors considered in this paper, it is more difficult to calculate 2D projections that approximate well the original distance distributions even with this SVD/minimization approach. This is because all scaled descriptors lie within the same range, and there are in general no dominant singular values. However, we showed that higher-accuracy projections can be obtained for these scaled datasets when the projection dimension is increased from two to ten. Though these higher-dimensional projections are not easily visualized, the compression of the dataset descriptors can be advantageous in further applications of the compound library analysis (e.g., diversity sampling). See ref 22 for results of a different scaling, an illustration of a 3D projection, and application of the mapping to similarity and diversity sampling.

When the intercompound distances in 2D approximate the original distance relationships well, the 2D projection offers a simple visualization tool for analyzing the compounds in a large database. The preliminary analysis offered here of the correspondence between chemical structure identity and relative position of the corresponding projection points indicates that the projection appears reasonable in terms of formal chemical similarity/diversity. Compounds that belong to the same cluster also generally belong to the same chemical and pharmacological class. Furthermore, chemically similar compounds tend to cluster together even closer within individual pharmacological groups. Compounds whose chemical structure differ are generally found farther away. Further clustering analysis is warranted, though real data for this purpose are difficult to obtain in practice because real databases are generally proprietary.

The projection accuracy is high for the unscaled data because two dominant singular values can be found corresponding to descriptors with large relative magnitudes. This is not the case for the scaled data. However, the comparison of projection maps for the scaled vs unscaled dataset ARTF indicates that, at least qualitatively, both projections provide a fairly analogous pattern of distribution of different pharmacological groups as discussed above. The ultimate goal of any pharmaceutical database analysis method is to establish and understand structure−activity relationships for the database compounds and, from this prospective, the necessity of descriptor scaling for the database mapping remains an area of future research.

These clustering analyses serve as a first step in the study of related combinatorial chemistry questions dealing with large chemical databases, and we hope to examine these possibilities in future work. Another important application of our approach deals with comparisons between different databases in terms of their similarity and diversity. Such comparisons are an essential part of compound acquisition strategies currently employed by many pharmaceutical companies.[19] Rapid and efficient projection of a proprietary database alongside an external database may help in quickly evaluating their relative similarity/diversity and selecting the most different external compounds for acquisition and testing. We emphasize that these analyses depend on the quality of the original descriptors, an area of research on its own.[7] We invite interested readers to contact us about experimenting with our projection software SIVER (singular values and error refinement).

ACKNOWLEDGMENT

We are very grateful to Dr. Yvonne Martin for providing binary descriptors for MAO compounds and Dr. Charles Reynolds for providing Molconn-Z descriptors for the PEST dataset. Support by the National Science Foundation (Grants ASC-9157582 and BIR 94-23827EQ) and the National

Institutes of Health (Grant R01 GM55164-01A2) is gratefully acknowledged. T.S. is an investigator of the Howard Hughes Medical Institute.

REFERENCES AND NOTES

(1) *Molconn-Z, version 3.1*; Hall Associates Consulting: Quincy, MD, 1998.
(2) *Molconn-X, version 2.0*; Hall Associates Consulting: Quincy, MD, 1995.
(3) *NAG Fortran Library, Mark 17*; NAG Inc.: Opus Place, Suite 200, Downers Grove, IL, 1995.
(4) Agrafiotis, D. K. A new method for analyzing protein sequence relationships based on Sammon maps. *Protein Sci.* **1997**, *6*, 287−293.
(5) Boyd, D. B. Rational drug design: Controlling the size of the haystack. *Mod. Drug. Discov.* **1998**, *1*, 41−47.
(6) Bakonyi, M.; Johnson, C. R. The Euclidean distance matrix completion problem. *SIAM J. Matrix Anal. Appl.* **1995**, *16*, 646−654.
(7) Brown, R. D.; Martin, Y. C. Information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sciences*, **1997**, *37*, 1−9.
(8) Crippen, G. M.; Havel, T. F. *Distance Geometry and Molecular Conformation*; Wiley, New York, 1988.
(9) Glunt, W.; Hayden, T. L.; Hong, S.; Wells, J. An alternating projection algorithm for computing the nearest Euclidean distance matrix. *SIAM J. Matrix Anal. Appl.* **1990**, *11*, 589−600.
(10) Golub, G. H.; Van Loan, C. F. *Matrix Computations*, 3rd ed.; John Hopkins University Press: Baltimore, MD, 1996.
(11) Gower, J. C. Properties of Euclidean and non-Euclidean distance matrices. *Linear Algebra Appl.* **1985**, *67*, 81−97.
(12) Hassan, M.; Bielawski, J. P.; Hempel, J. C.; Waldman, M. Optimization and visualization of molecular diversity and combinatorial libraries. *Mol. Divers.* **1996**, *2*, 64−74.
(13) Kohonen, T. *Self-Organizing Maps, Springer Series in Information Sciences*; Springer: Berlin, Heidelberg, New York, 1997; Vol. 30.
(14) *MACCS-II*; Molecular Design Ltd.: 14600 Catalina St., San Leandro, CA 94577, 1998.
(15) Moré, J. J.; Wu, Z. *Distance geometry optimization for protein structures*; MCS-P628-1296; Argonne National Laboratory: Argonne, IL, 1997.
(16) Pinou, T., Schlick, T., Li, B.; Dowling, H. G. Addition of Darwin's third dimension to phyletic trees. *J. Theor. Biol.* **1996**, *182*, 505−512.
(17) Robinson, D. D.; Barlow, T. W.; Richard, W. G. Reduced dimensional representations of molecular structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 939−942.
(18) Schlick, T.; Fogelson, A. TNPACK—A truncated Newton minimization package for large-scale problems: I. Algorithm and usage. *ACM Trans. Math. Softw.* **1992**, *14*, 46−70.
(19) Shemetulskis, N.; Dunbar, J., Jr.; Dunbar, B.; Moreland, D.; Humblet, C. Enhancing the diversity of a corporate database using chemical database clustering and analysis. *Comput.-Aided Mol. Des.* **1995**, *9*, 407−416.
(20) Xie, D.; Schlick, T. Efficient implementation of the truncated-Newton algorithm for large-scale chemistry applications. *SIAM J. Optim.* **1999**, in press.
(21) Xie, D.; Schlick, T. Remark on Algorithm 702—the updated truncated Newton minimization package. *ACM Trans. Math. Softw.* **1999**, *25*, 108−122.
(22) Xie, D.; Schlick, T. Visualization of chemical databases using the singular value decomposition and truncated-Newton minimization. Preprint, submitted to *Optimization in Computational Chemistry and Molecular Biology*; Floudas, A., Pardalos, P. M., Eds.; Kluwer Academic Publishers B. V.: Dordrecht, The Netherlands, in press.
(23) Willett, P. Structural similarity measures for database searching. In *Encyclopedia of Computational Chemistry*; Schleyer, P. v. R., Editor-in-Chief; Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III, Eds.; John Wiley & Sons: West Sussex, U.K. 1998; Vol. 4, pp 2748−2756.

CI990333J