SUPPLEMENTARY MATERIALS: ADDITIONAL TABLES AND FIGURES

#Pivots	γ	Recall	
		top 10	top 100
15	1	13.0%	2.5%
	2	13.5%	3.6%
	5	14.4%	6.2%
	10	15.6%	9.7%
	30	19.4%	20.0%
700	1	12.0%	1.0%
	2	12.1%	1.3%
	5	12.2%	2.2%
	10	12.3%	3.6%
	30	13.0%	6.1%

Table S-1. *Embedding quality of a non-linear mapping network method benchmarked by recall rates of similarity searches.* An embedding algorithm known as *non-linear mapping networks* (Agrafiotis et al., 2001) was implemented using the `nnet` package in R 2.10. The input to the network is a set of similarity keys, which are obtained by comparing a compound to a set of pivot compounds. The embedding experiments were performed with the NCI data set using different numbers of pivots and 3000 randomly selected compounds to train the network. Each compound was embedded into a 100-D vectors. With these embedding results, 1000 random similarity searches were carried out using different relaxation ratios γ. The table lists the average recall rates that were achieved in these tests. Both recall rate and relaxation ratio are defined in Section 3.5.

#Dimensions	#Steps	Average Embedding Time (in second per compound)	Recall	
			top 10	top 100
2	1,000,000	0.012	10.3%	2.6%
	5,000,000	0.054	10.2%	2.9%
	10,000,000	0.099	10.4%	3.0%
	32,000,000	0.286	10.2%	3.3%
	50,000,000	0.432	10.3%	4.4%
120	1,000,000	0.012	13.3%	21.4%
	5,000,000	0.063	55.5%	80.0%
	10,000,000	0.116	77.8%	91.2%
	20,000,000	0.179	88.7%	95.4%
	32,000,000	0.340	94.2%	97.6%
	50,000,000	0.535	97.0%	98.8%
Modified MDS (120D)		0.259	97.9%	96.4%
Modified MDS (120D)		0.146	95.9%	95.0%

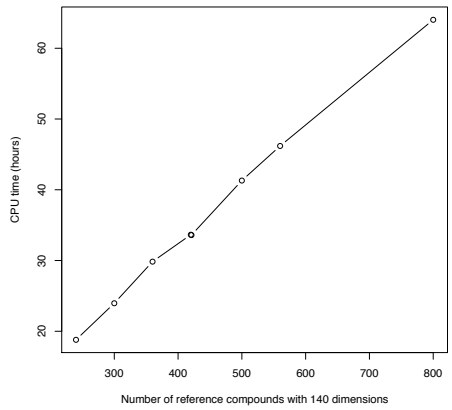
Table S-2. *Embedding efficiency and accuracy of SPE-based embedding method compared to EI's modified MDS method.* The table gives the average per-compound embedding times and recall rates for the NCI data sets when using the SPE algorithm or EI's modified MDS method. Embedding with SPE was carried out using different number of output dimensions and different number of steps (first and second column). Recall rates, as defined in Section 3.5, are given for searches of the 10 and 100 most similar neighbors of 1000 randomly selected query compounds. Relaxation ratio is set to 30 in all tests. All experimental parameters for the SPE method were chosen according the recommendations from Agrafiotis and Xu (2002). The corresponding results from our modified MDS algorithm are given in the last two rows.

Data set	NCI	PubChem Subset	PubChem Compound
Size	260,071	2,288,680	19,629,027
Embedding time			
Total (hours)	18.72	179.24	1543.83
Per-compound (seconds)	0.259	0.282	0.284

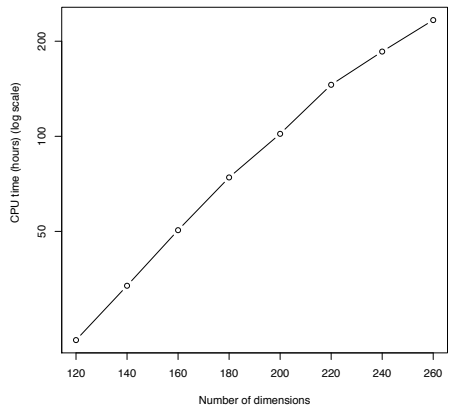
Table S-3. *Processing time for embedding.* The table gives the time required for embedding each compound library as well as the average time required per compound. All experiments were performed on the same hardware using the same parameters (R=300 and D=120).

Parameters	Average Embedding Time (in second per compound)	Recall	
		top 10	top 100
FACTR=10^{10}, PGTOL=0.001	0.526	98.0%	97.0%
FACTR=10^{12}, PGTOL=0.010	0.259	97.9%	96.4%
FACTR=10^{13}, PGTOL=0.040	0.146	95.9%	95.0%
FACTR=10^{16}, PGTOL=0.100	0.035	86.9%	85.7%

Table S-4. *Processing times for the embedding step using different optimization parameters.* The table gives the average per-compound embedding times for the NCI data set using the indicated optimization parameters for the L-BGFS-B algorithm. All experiments were performed on the same hardware using the same parameters (retrieval of top 100 compounds, R=300, D=120 and γ=30).



(a)



(b)

Fig. S-1: *Embedding time and parameters*. The graphs show the impact of the sizes of reference compound set R (a) and the number of dimensions D (b) on the total CPU time required for embedding the compounds from the NCI data set in a high-dimensional Euclidean space. In graph (b) R was set to three times the value of D.

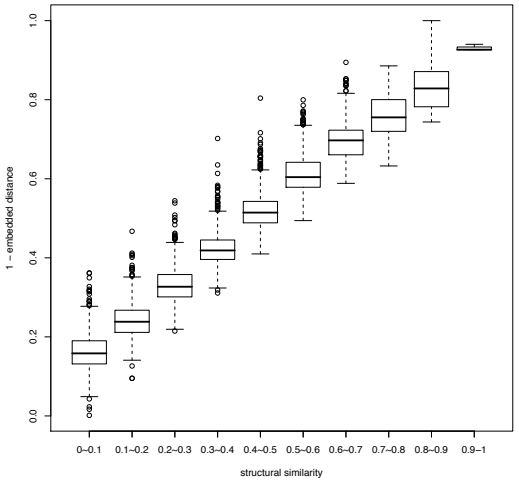
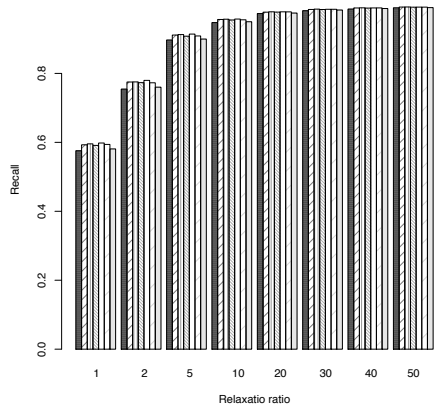
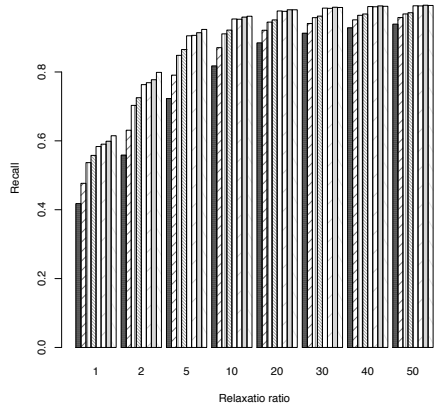


Fig. S-2: *Embedding accuracy.* The box plots compare the compound-to-compound distances obtained from the embedding method with the corresponding Tanimoto similarities for ten intervals ranging from 0-1.



(a)



(b)

Fig. S-3: *Recall rates of the EI-Search method.* Panel (a) provides the recall rates of EI-Search using different relaxation ratios γ and numbers of reference compounds R. Both recall rate and relaxation ratio is defined in Section 3.5. Panel (b) shows the same data for variable γ and D values. In panel (a) the chosen R values for each γ were: 240, 300, 360, 420, 500, 560, and 800, and 260 (from left to right). In panel (b), the D values were: 40, 60, 80, 100, 120, 140, 160 and 180 .