Unbiased Chemical Space Visualization

Breton M. Saunders
Inpharmatica Cambridge
b.saunders@inpharmatica.co.uk

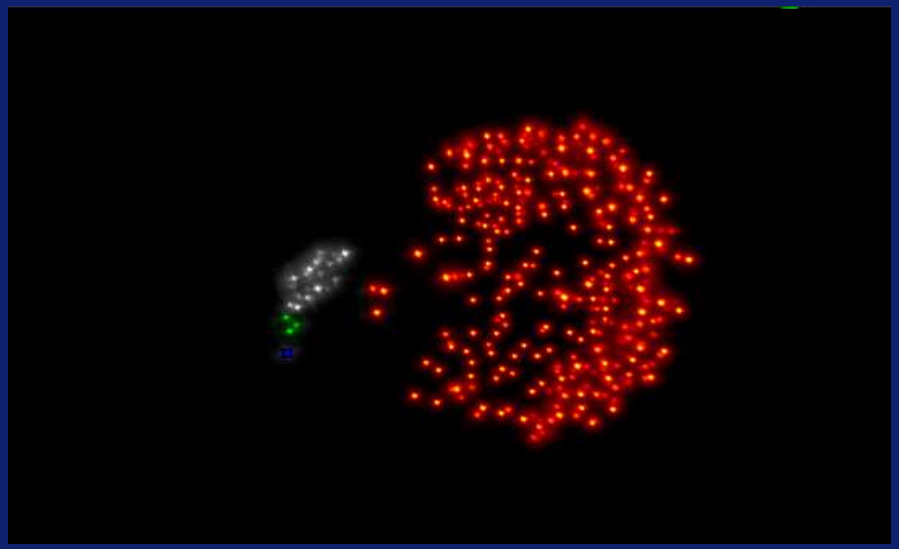
History

- Inpharmatica (Cambridge Science Park)
  - Was Camitro/ArQule
    - Predict ADME Properties for internal projects only
- Now as Inpharmatica
  - Analyse customer and internal compounds in probabilistic framework
  - ADME Lead Optimization Consultancy
  - *In vitro* ADME screening, Assay development
    - Custom (per project) predictive model development
  - http://www.inpharmatica.co.uk

inpharmatica

"Chemical Space"

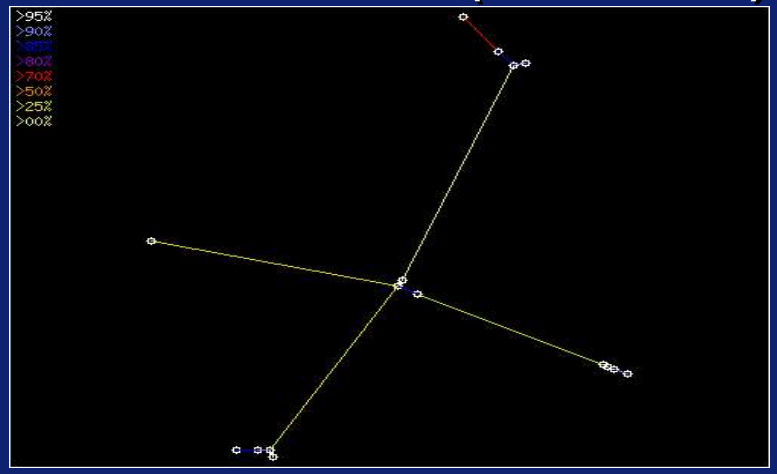
- What is it?
  - A poorly defined notion
  - Great for executives to get excited over
  - Our definition: "A space in which distance between points approximates compound similarity"



inpharmatica

Other Approaches

- Hierarchical
  - Diversity Map, Bernard Rohde (Novartis)

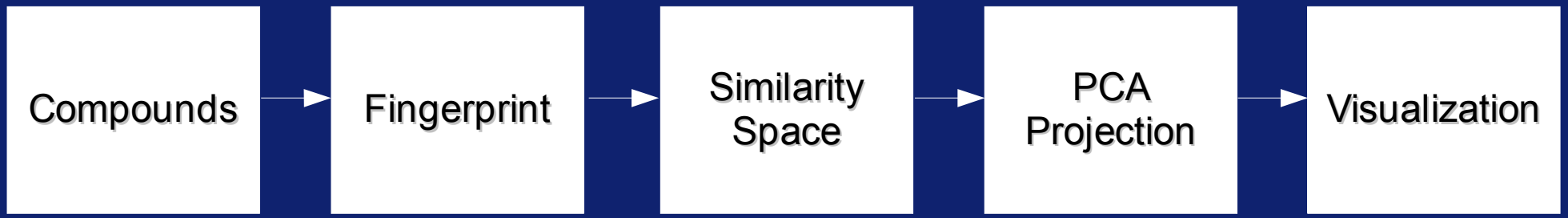


- Unbiased
  - All compounds have equal importance
  - Minimisation approaches ($d_{ij} = 1.0 - s_{ij}$)
    - Conjugate Gradient/Simulated Annealing
    - Poor performing (in performance and results)
  - Kohonen maps

inpharmatica

Our Approach

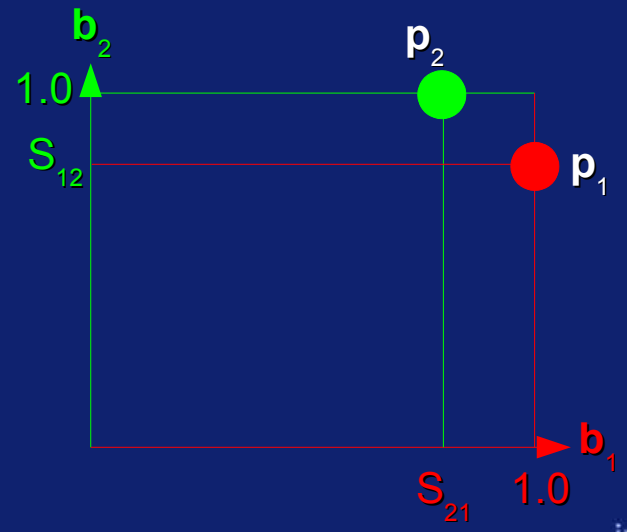
- Direct projection of the similarity matrix using PCA
- Each compound acts as a basis direction for a "Similarity Space"
- Project the similarity space to $\mathbf{R}^2$ or $\mathbf{R}^3$ for visualization



inpharmatica

Similarity Space Set-up

- Similarity Metric
  - Daylight fingerprint: 1024 bits
  - Chain lengths between 0 and 10 bonds
  - Tanimoto similarity metric:
    $s_{ij} = |\mathbf{B}_i \wedge \mathbf{B}_j| / (|\mathbf{B}_i| + |\mathbf{B}_j| - |\mathbf{B}_i \wedge \mathbf{B}_j|)$
- Space setup
  - $\mathbf{p}_i = [s_{i1}, s_{i2}, \ldots s_{ij}, \ldots s_{iN}]$
  - $\mathbf{p}_1 = [1.0, s_{12}]$
  - $\mathbf{p}_2 = [s_{21}, 1.0]$

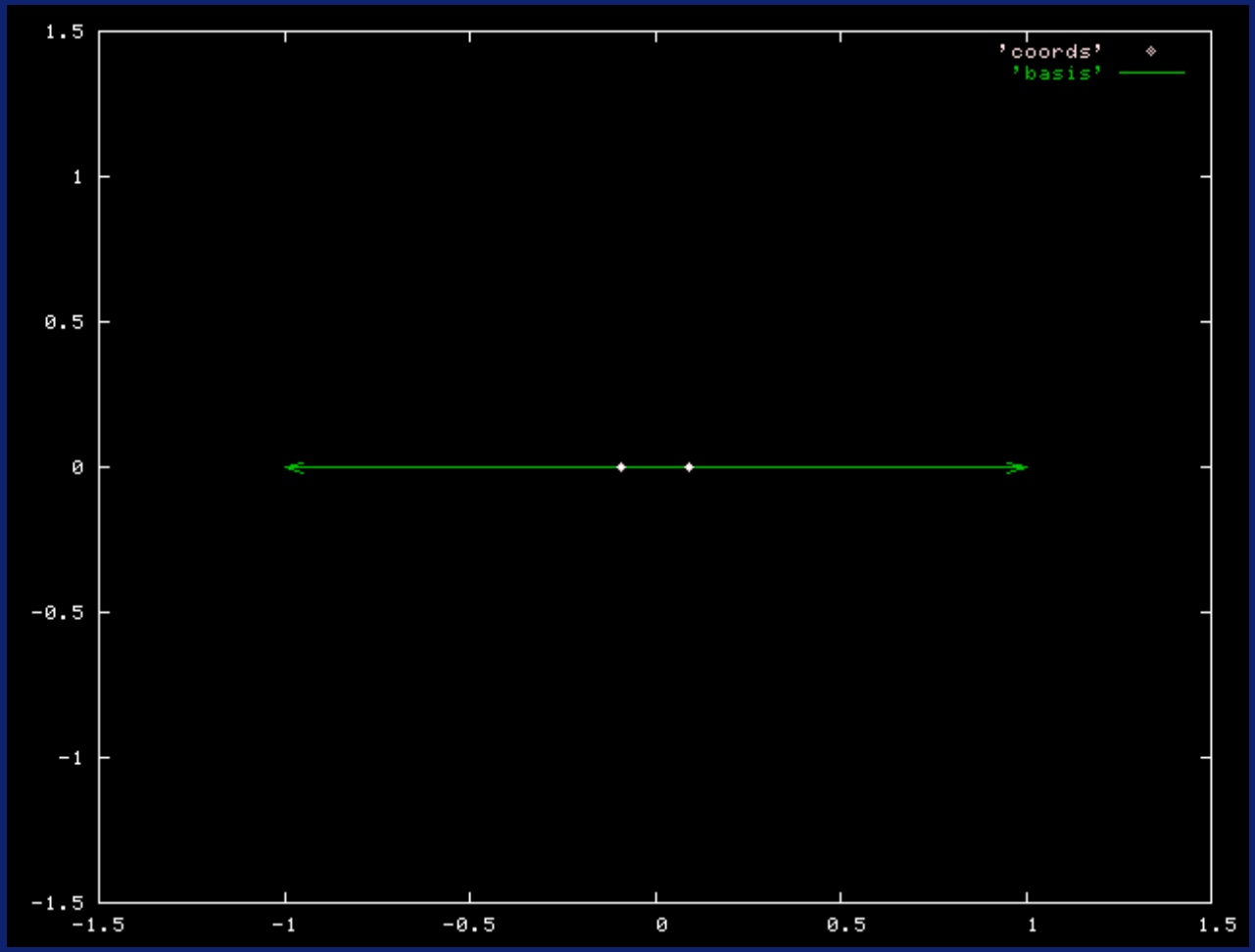


inpharmatica

Projection to low dimension space

- People are not good at thinking in $\mathbb{R}^N$ where N>3!
- Projection to low dimension space should:
  - Remove redundant information
  - Maximise variance in the source data
    - "Show me the large differences in the source data and hide the small ones"
- The least-squares projection (aka PCA) accomplishes these

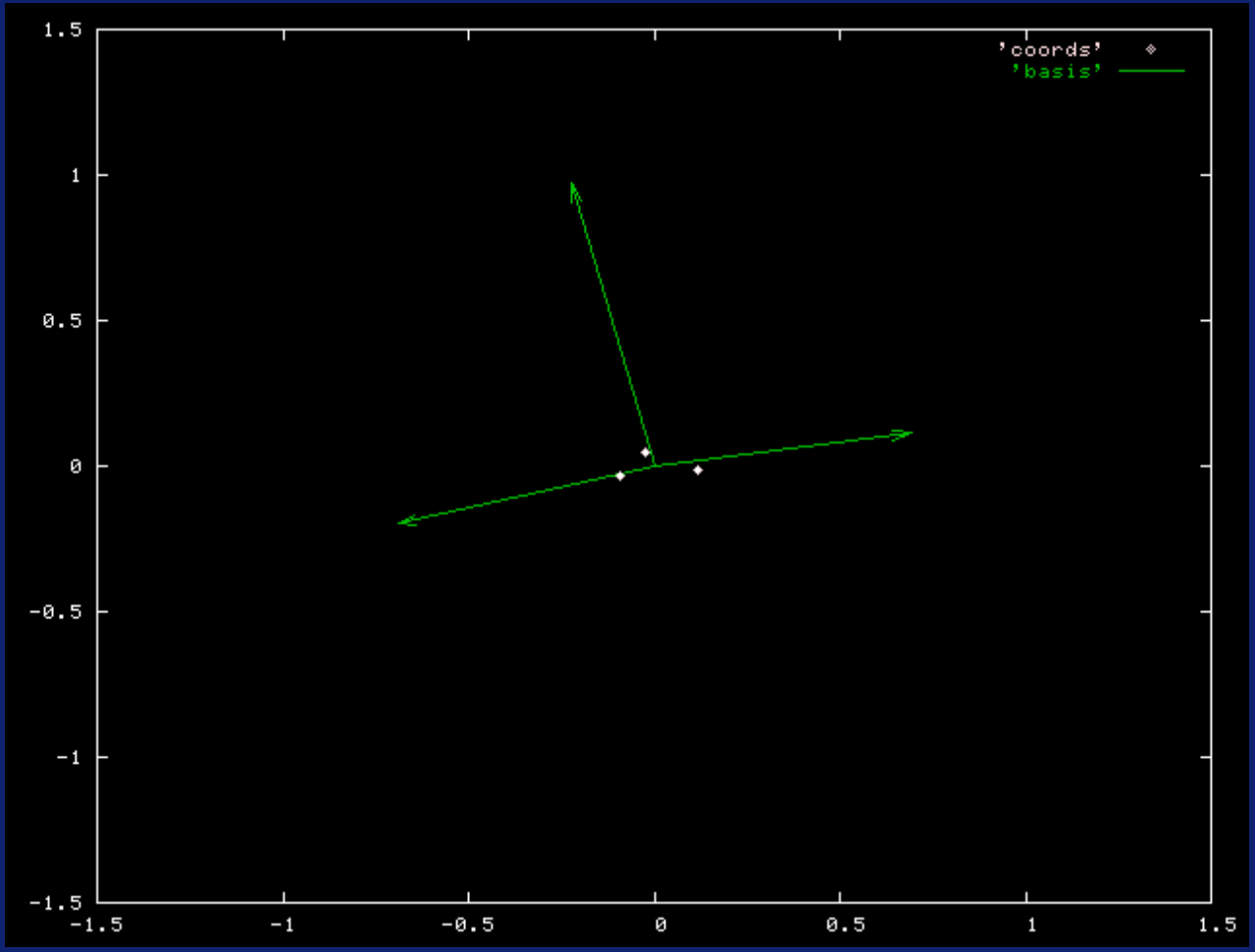
Graphical Example



Max d_{ij} = 0.093

inpharmatica

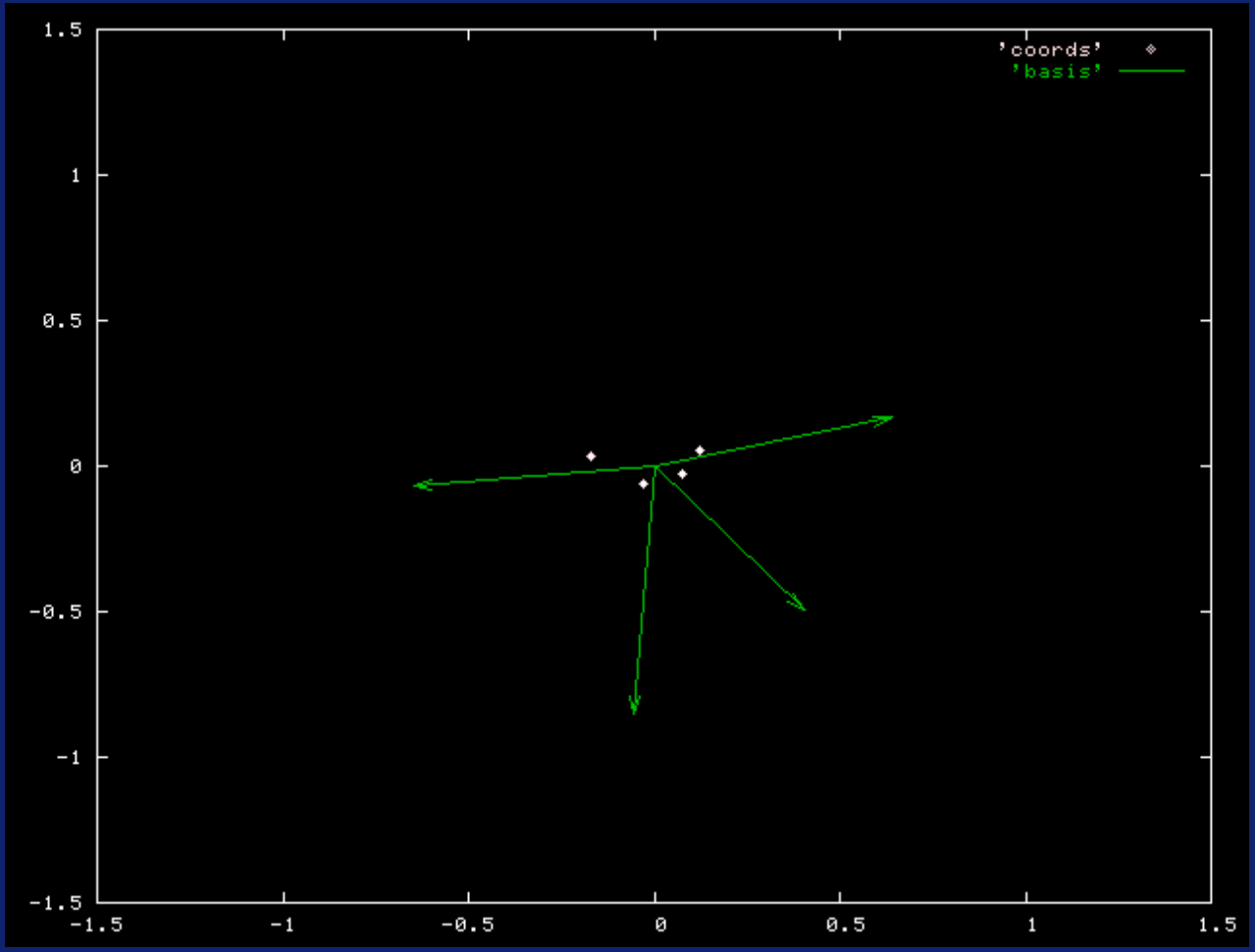
Graphical Example



Max d_{ij} = 0.149

inpharmatica

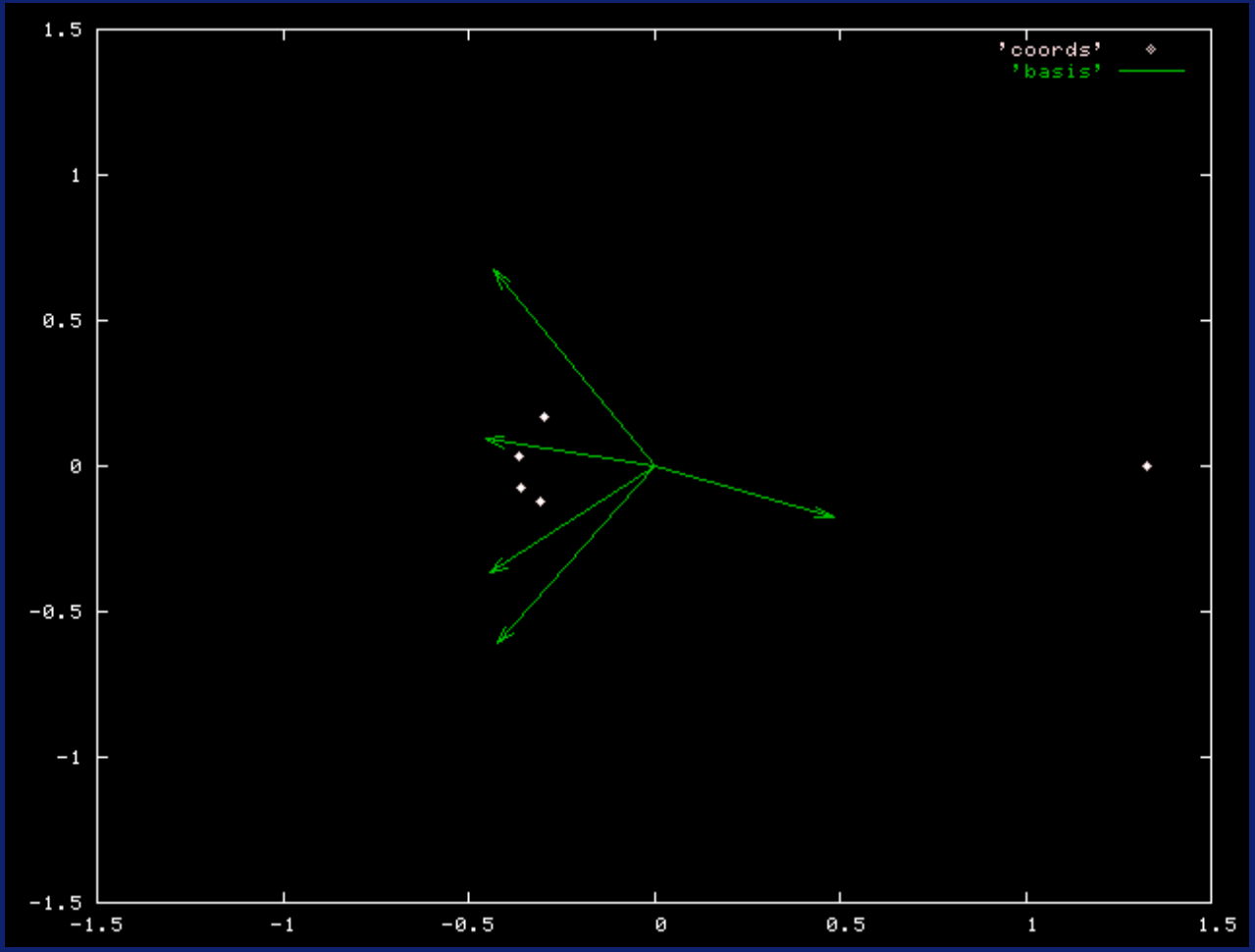
Graphical Example



Max d_{ij} = 0.195

inpharmatica

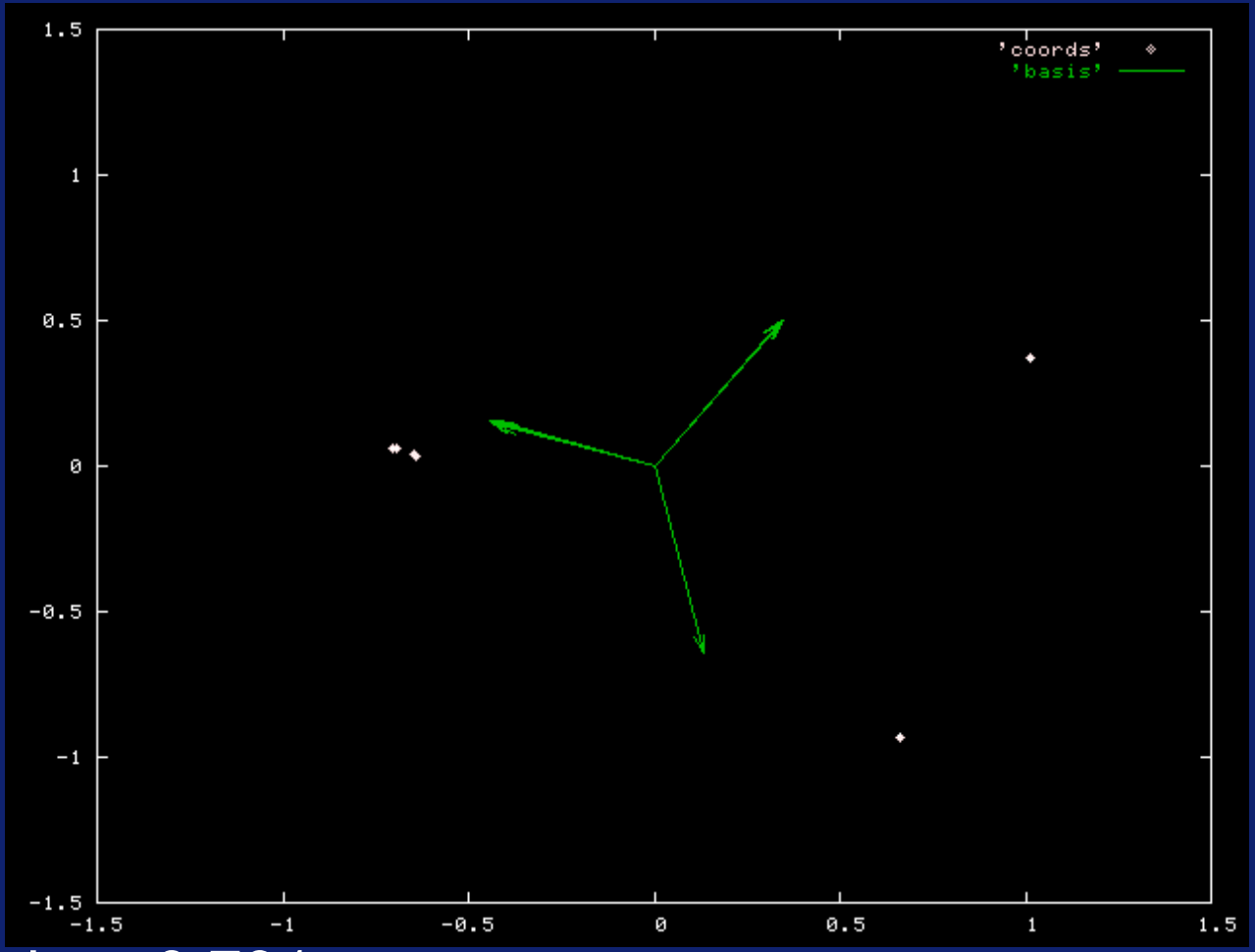
Graphical Example



Max d_{ij} = 0.83

inpharmatica

Graphical Example



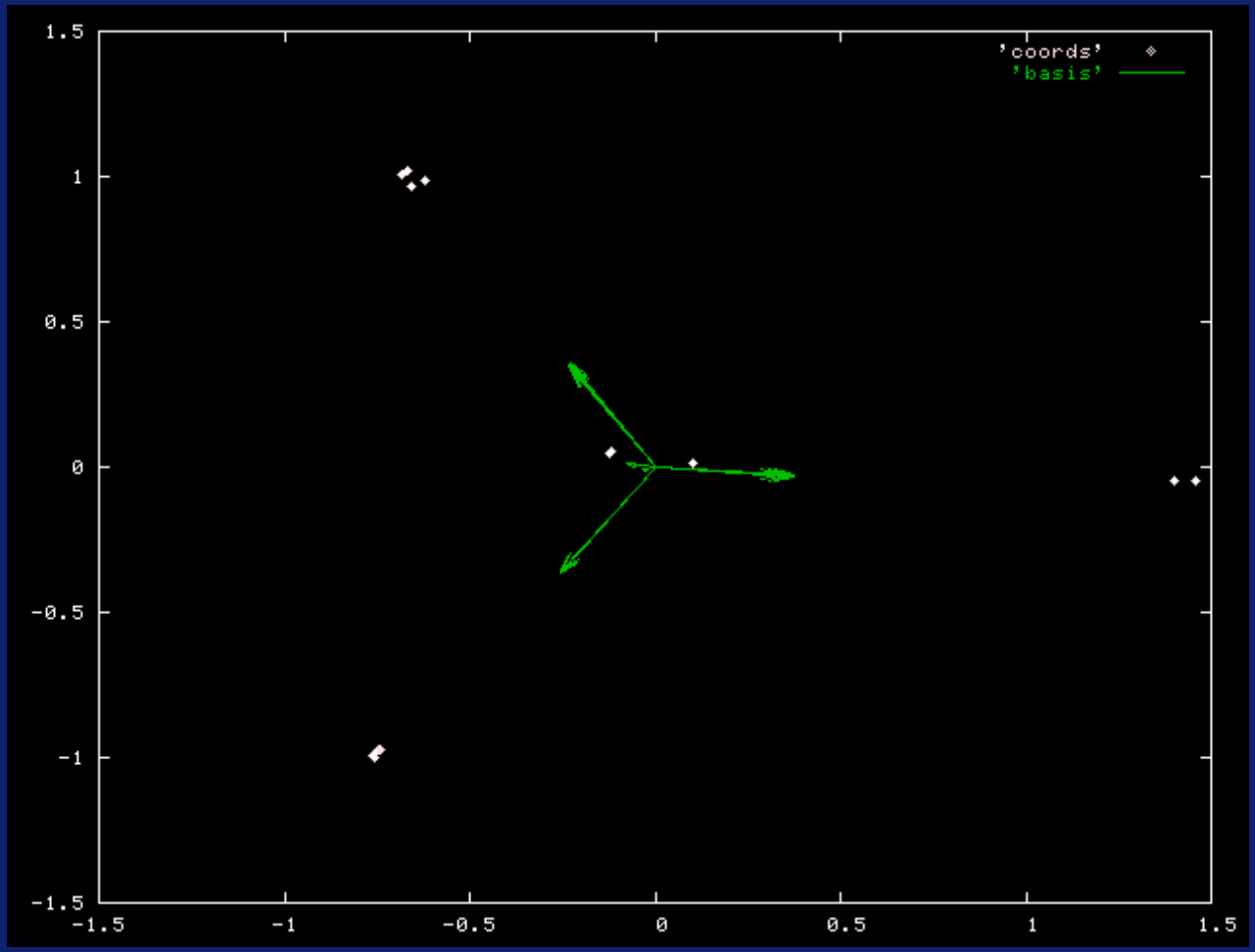
$d_{65} = 0.791$

Min $d_{5[1,2,3,4]} = 0.8$

Min $d_{6[1,2,3,4]} = 0.75$

inpharmatica

Graphical Example



inpharmatica

PCA Implementation

- Normally

  [**u**,**s**,**v**] = svd(cov(**y**)) (**y** = mean centered observations)
  W = [**v**(:,1), **v**(:,2), ...] (the eigenvectors)
  x = **yW** (**x** = coordinates in latent space)

- Complexity
  - SVD is $O(N^3)$ in runtime
  - COV is $O(N^2)$ in memory
  - 8000 compounds: terminated after ~3 days on 1.6Ghz Athlon w/1Gb ram

- Search for a better PCA algorithm!

inpharmatica

Expectation Maximization for PCA

- Solution: Use Roweis' EM algorithm
  - http://www.cs.toronto.edu/~roweis/
- EMPCA is O(knp)
  - k = # of eigenvectors
  - n = # of observations
  - p = # of observed dimensions
  - $O(2n^2)$ in our case
- Memory requirement is O(n)
  - PCs are updated using rank-one update

EMPCA Algorithm Overview

- Treat PCA as a linear Gaussian model:

  $\mathbf{y} = \mathbf{C}\mathbf{x} + \mathbf{v}$

  $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})$ (latent coordinates)

  $\mathbf{v} \sim \mathcal{N}(0, \mathbf{R})$ (noise)

  $\mathbf{y} \sim \mathcal{N}(0, \mathbf{C}\mathbf{C}^T + \mathbf{R})$ (distribution of observed data)

- Assume no noise (least-squares condition)

  Expected latent variables (e-step): $\mathbf{x} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{y}$

  Learning step (m-step): $\mathbf{C}^{new} = \mathbf{y}\mathbf{x}^T(\mathbf{x}\mathbf{x}^T)^{-1}$

- The EM steps are iterated to convergence

inpharmatica

Implementation Key Points

- C++ / MMX asm code for $|\mathbf{B}_i \hat{} \mathbf{B}_j|$ calculation
- Matrix Template Library used for math
  - A poor choice in hindsight
- 8000 compound set: 25 min CPU time
- ~40% CPU time spent in $|\mathbf{B}_i \hat{} \mathbf{B}_j|$ calculation
  - Ideally, this would be closer to 90%
  - MTL iterators use excessive CPU time

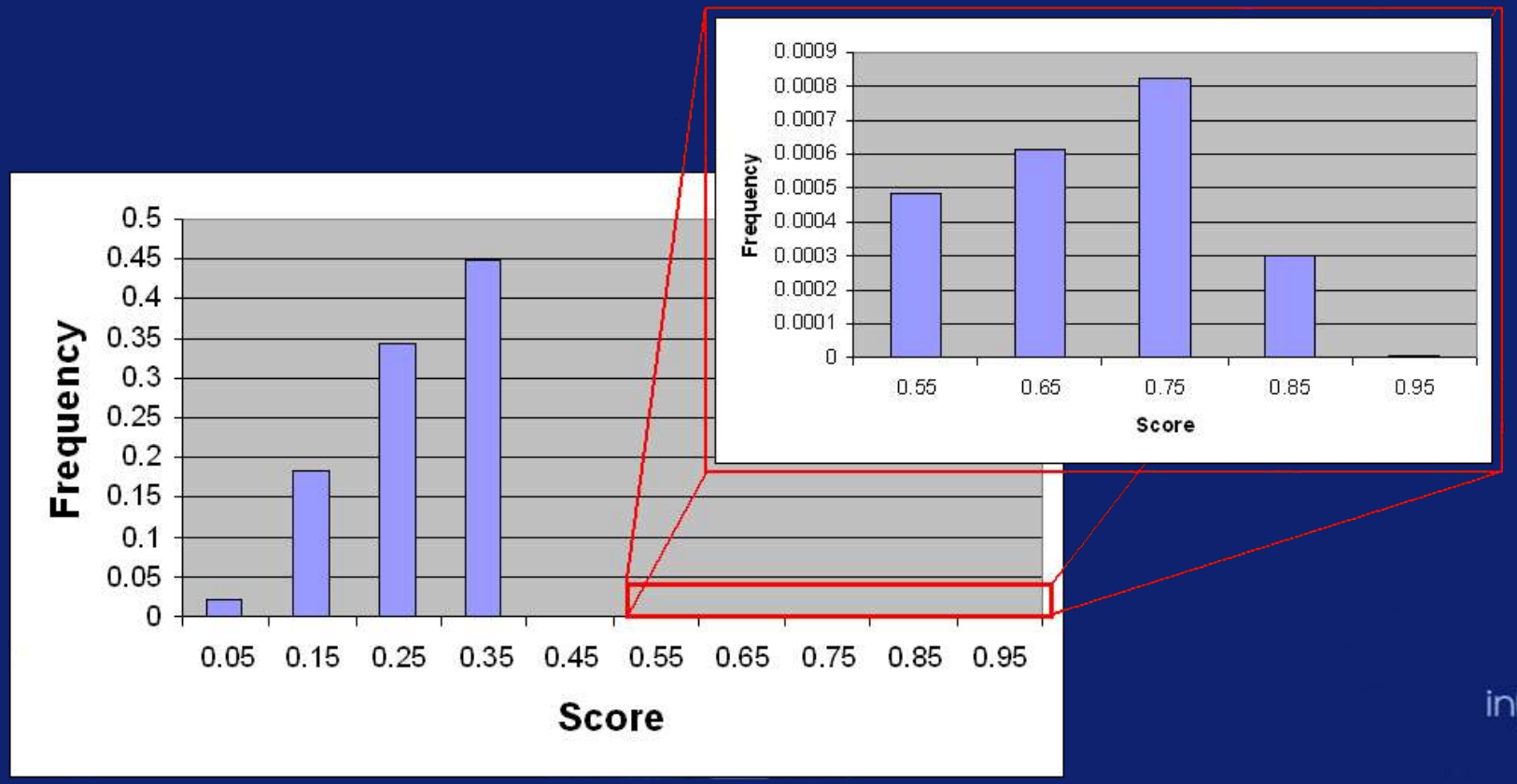
Usage at Inpharmatica

- Process
  - Predict ADME properties and uncertainties
  - Perform probabilistic scoring
  - Generate an “optimal” sub-set selection for:
    - Combinatorial synthesis
    - Non-combinatorial selection
  - Visualize the selected subset and contrast with the full compound set
    - Representative sub-set selection to reduce data size

inpharmatica

Combinatorial Set Picking

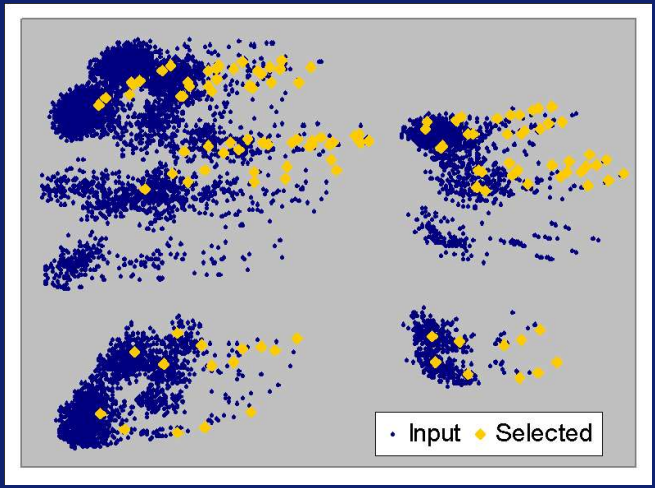
- Customer Compound Set: 350k cmpds
  - 3D eunmeration space
  - Select a 5x5x5 subset for further work

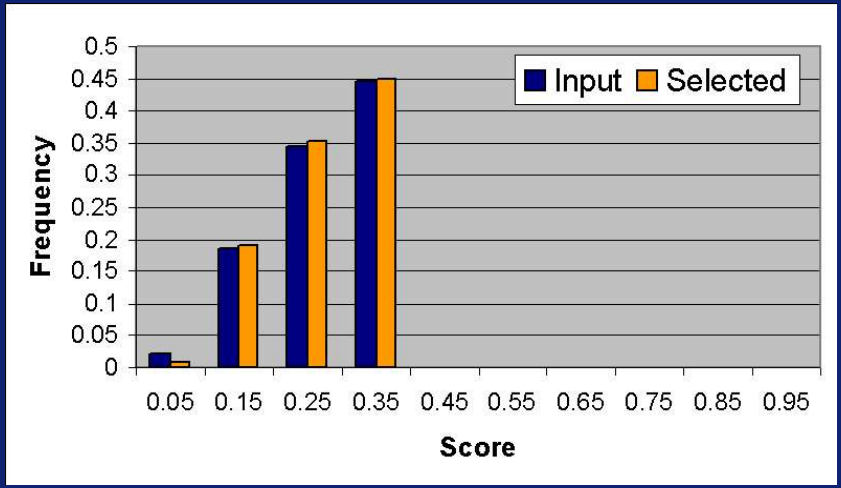


inpharmatica

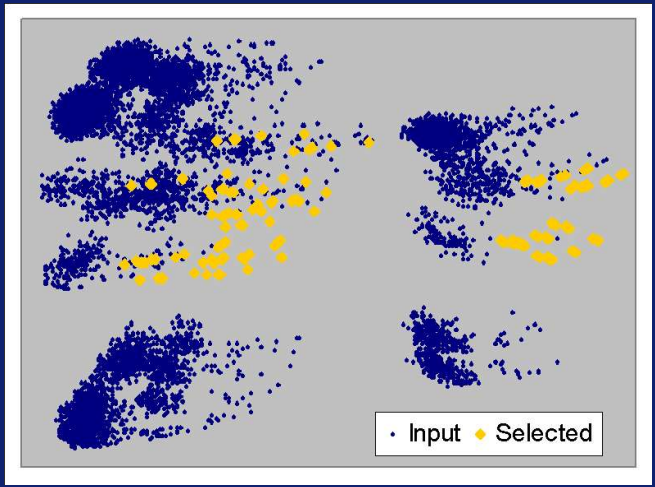
Combinatorial Selection Example

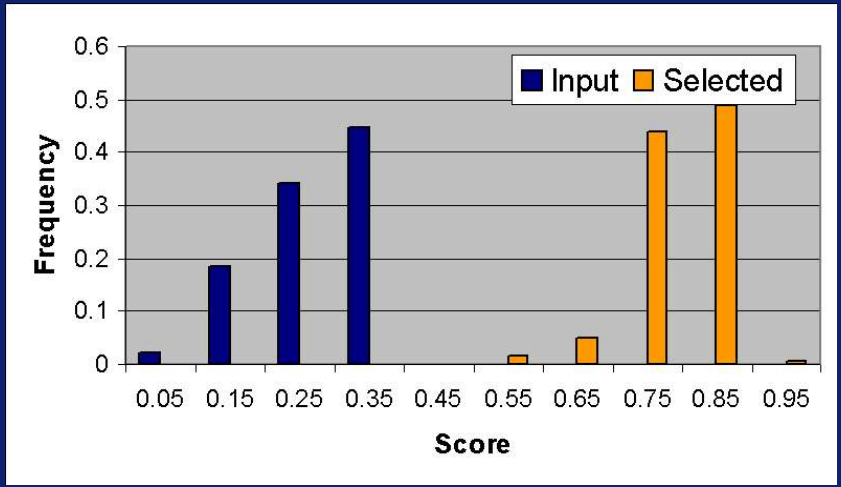
Diverse Selection 125 cmpds, N=10,125





Balanced Diversity / Score Selection 125 cmpds, N=10,125





inpharmatica

Marketing Example

inpharmatica

Future Work

- Better similarity space model
- Projection Pursuit and ICA
- Preconditioning using clustering
- Open Source Code?

Conclusions

- Objective:
  - Create an unbiased chemical space visualization
- Implementation:
  - PCA projection
  - Fast implementation using Roweis' EM algorithm
  - Brief implementation details
- Result:
  - The technique is useful for visualizing compound set selection and “chemical space”

inpharmatica

Acknowledgments

Mathew Segall
Ed Champness
Darko Butina

inpharmatica

Literature

EM algorithms for PCA and Sensible PCA

Roweis, S (1997), Technical Report CNS-TR-97-02
California Institute of Technology
http://citeseer.nj.nec.com/roweis97em.html

Maximum likelihood from incomplete data via the EM algorithm

Journal of the Royal Statistical Society Series B,
vol 39, no. 1, pp. 1-38, Nov 1977

Data Sets

- Compound set used in "Graphical Example" example from Bernhard Rhode's online example at http://www.daylight.com/cgi-bin/novartis/diversity/DiversityMap.cgi
- Smiles

```
Cc1ccc2c(Br)cc(Br)c(O)c2n1 s1
CCc1ccc2c(Br)cc(Br)c(O)c2n1 s2
CCCc1ccc2c(Br)cc(Br)c(O)c2n1 s3
CCCCc1ccc2c(Br)cc(Br)c(O)c2n1 s4
CCCCc1cc(Cl)c(cc1S(=O)(=O)N)S(=O)(=O)N s5
Oc1ccc(cc1O)C2CNCc3sccc23 s6
COc1ccc(cc1O)C2CNCc3sccc23 s7
CCOc1ccc(cc1O)C2CNCc3sccc23 s8
CCCOc1ccc(cc1O)C2CNCc3sccc23 s9
CN(C)CCCNCC(O)COc1ccccc1C(=O)CCc2ccccc2 s10
CCC(C)(C)NCC(O)COc1ccccc1C(=O)CCc2ccccc2 s11
CCCCC(C)(C)NCC(O)COc1ccccc1C(=O)CCc2ccccc2 s12
OCC(O)C(O)C(O)C=N s13
COCC(O)C(O)C(O)C=N s14
CCOCC(O)C(O)C(O)C=N s15
CCOCC(O)C(O)C(O)C=N s16
```

- Similarity Matrix

```
1 0.907895 0.851852 0.805447 0.170807 0.223629 0.22268 0.225152 0.222664 0.225989 0.223496 0.226257 0.0376569 0.0445344 0.0433071 0.0503876
0.907895 1 0.938272 0.88716 0.18806 0.236626 0.235412 0.237624 0.237354 0.246575 0.244444 0.246612 0.0466926 0.0528302 0.0514706 0.057971
0.851852 0.938272 1 0.945525 0.200581 0.249493 0.248016 0.25 0.254335 0.263441 0.26158 0.263298 0.0518519 0.057554 0.0561404 0.0622837
0.805447 0.88716 0.945525 1 0.213068 0.25 0.248544 0.250478 0.254717 0.260417 0.261905 0.263566 0.0530035 0.0584192 0.057047 0.0629139
0.170807 0.18806 0.200581 0.213068 1 0.209354 0.211329 0.208955 0.214286 0.232919 0.242038 0.248447 0.116402 0.121827 0.123153 0.119617
0.223629 0.236626 0.249493 0.25 0.209354 1 0.966321 0.939547 0.914216 0.25 0.251055 0.247423 0.0588235 0.0653266 0.0694789 0.0684597
0.22268 0.235412 0.248016 0.248544 0.211329 0.966321 1 0.972292 0.946078 0.256148 0.257261 0.256098 0.0621891 0.0684597 0.0724638 0.0714286
0.225152 0.237624 0.25 0.250478 0.208955 0.939547 0.972292 1 0.973039 0.258065 0.259184 0.258 0.0605327 0.0666667 0.0705882 0.0696056
0.222664 0.237354 0.254335 0.254717 0.214286 0.914216 0.946078 0.973039 1 0.272545 0.273834 0.272366 0.063981 0.0699301 0.0737327 0.0727273
0.225989 0.246575 0.263441 0.260417 0.232919 0.25 0.256148 0.258065 0.272545 1 0.935065 0.888889 0.121339 0.125506 0.135458 0.141176
0.223496 0.244444 0.26158 0.261905 0.242038 0.251055 0.257261 0.259184 0.273834 0.935065 1 0.948276 0.125 0.129167 0.134694 0.131474
0.226257 0.246612 0.263298 0.263566 0.248447 0.247423 0.256098 0.258 0.272366 0.888889 0.948276 1 0.123457 0.12749 0.137255 0.1341
0.0376569 0.0466926 0.0518519 0.0530035 0.116402 0.0588235 0.0621891 0.0605327 0.063981 0.121339 0.125 0.123457 1 0.803922 0.706897 0.640625
0.0445344 0.0528302 0.057554 0.0584192 0.121827 0.0653266 0.0684597 0.0666667 0.0699301 0.125506 0.129167 0.12749 0.803922 1 0.87931 0.796875
0.0433071 0.0514706 0.0561404 0.057047 0.123153 0.0694789 0.0724638 0.0705882 0.0737327 0.135458 0.134694 0.137255 0.706897 0.87931 1 0.90625
0.0503876 0.057971 0.0622837 0.0629139 0.119617 0.0684597 0.0714286 0.0696056 0.0727273 0.141176 0.131474 0.1341 0.640625 0.796875 0.90625 1
```

inpharmatica