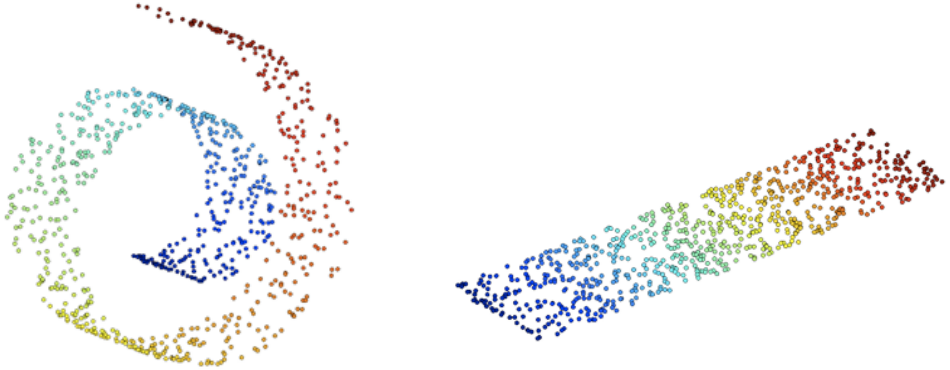
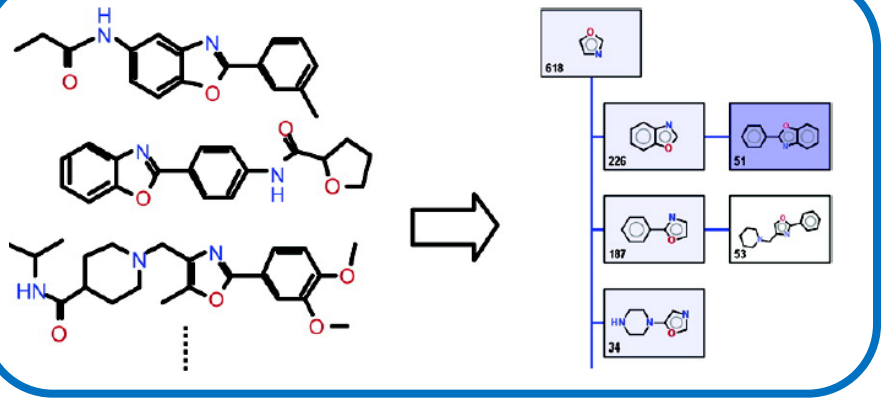
Введение в химическую информатику

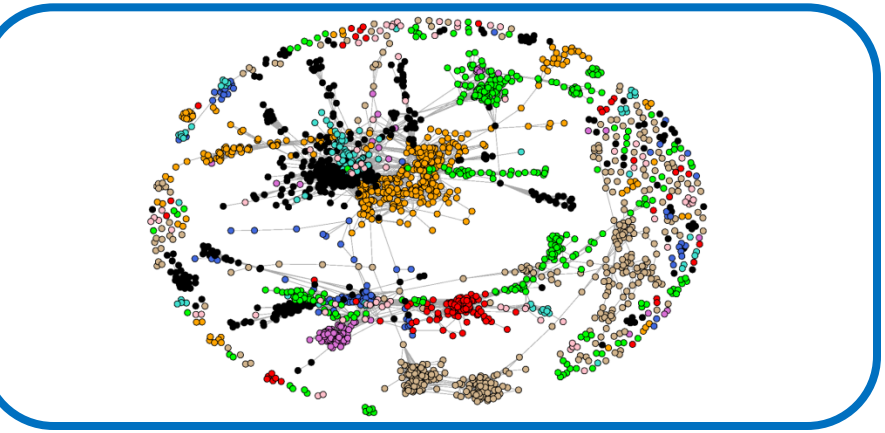
Лекция 10



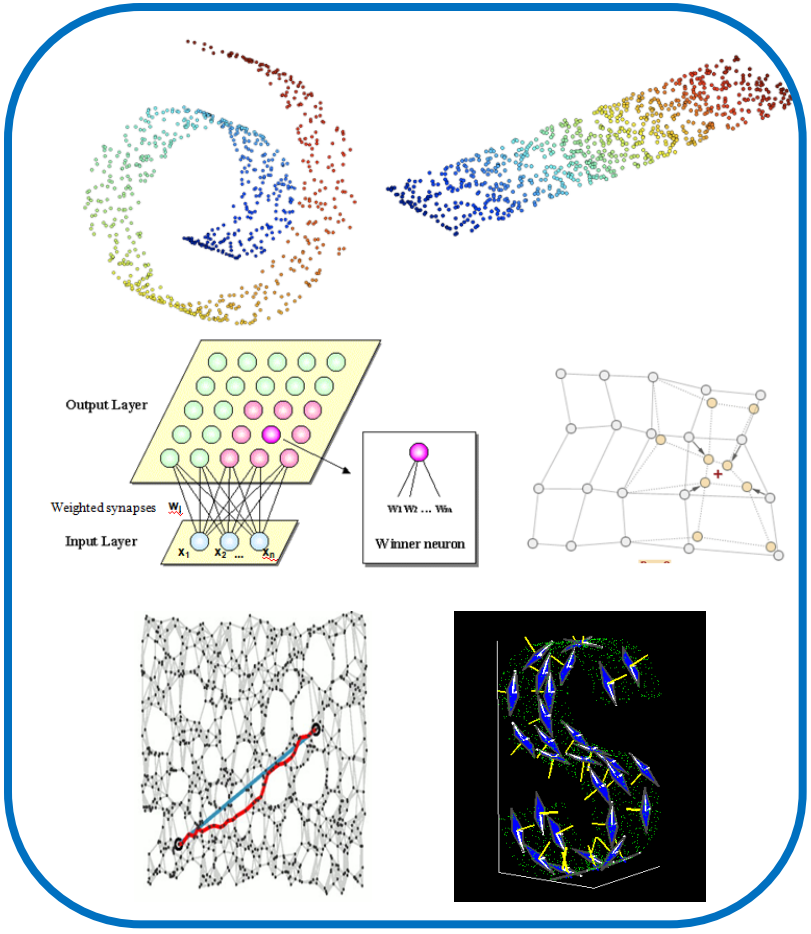
МЕТОДЫ ПРЕДСТАВЛЕНИЯ ХИМИЧЕСКОГО ПРОСТРАНСТВА



Методы, основанные на молекулярных графах



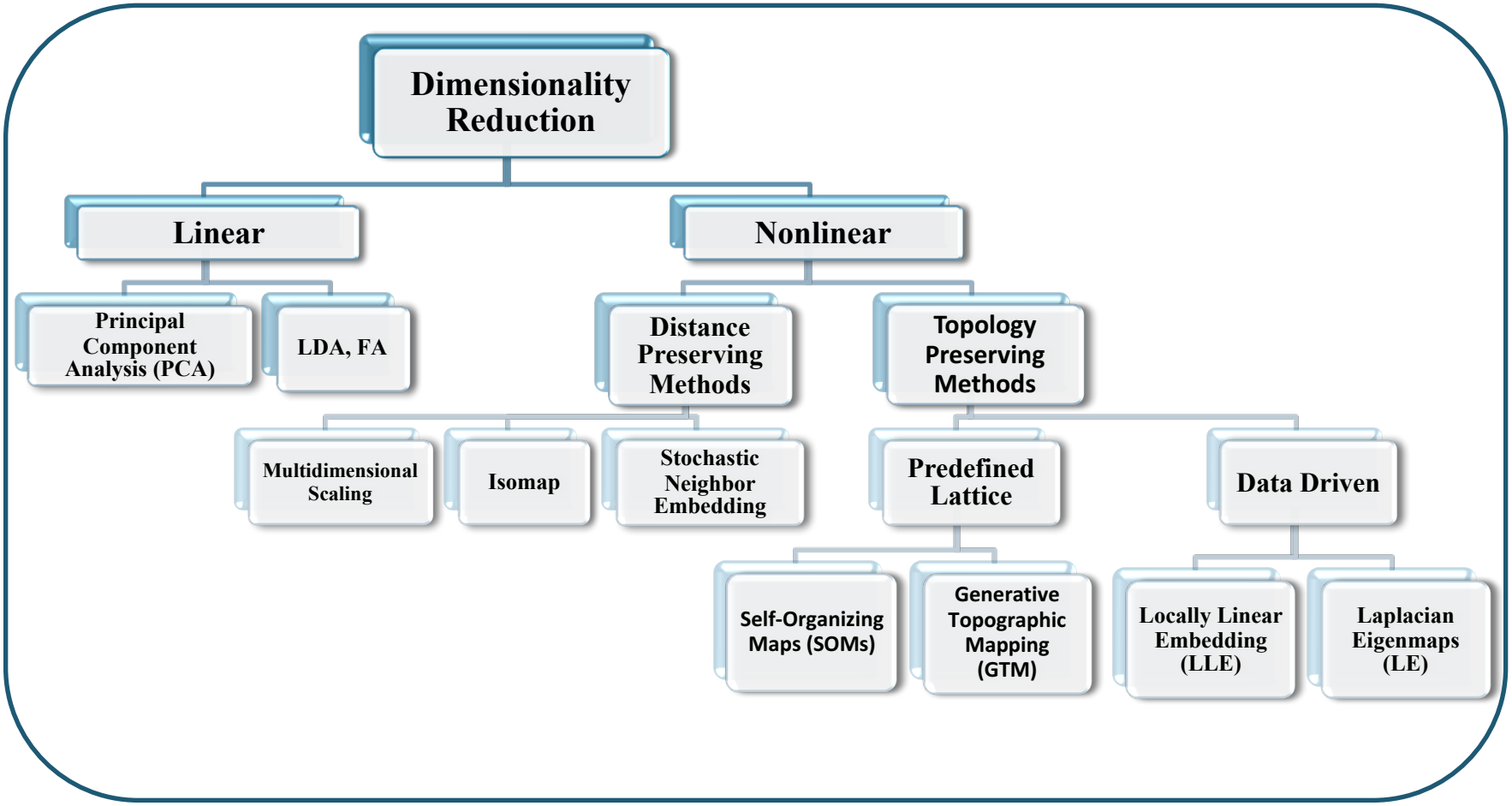
Сетевое представление химического пространства данных



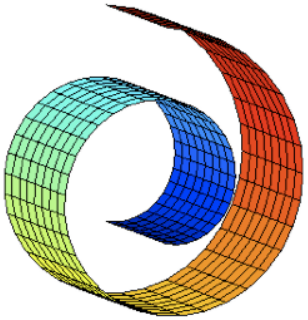
Методы, основанные на дескрипторном представлении данных

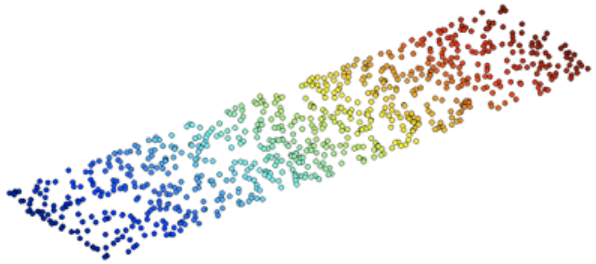
МЕТОДЫ, ОСНОВАННЫЕ НА ДЕСКРИПТОРНОМ ОПИСАНИИ ДАННЫХ (МЕТОДЫ ПОНИЖЕНИЯ РАЗМЕРНОСТИ ДАННЫХ)

- Сжатие химического пространства к размерности 2 или 3 с минимальными потерями информации
- Сохранение принципа топологии (соседние в многомерном пространстве объекты должны оставаться таковыми в пространстве пониженной размерности)
- Сохранение расстояний между объектами (степень подобия (сходства))

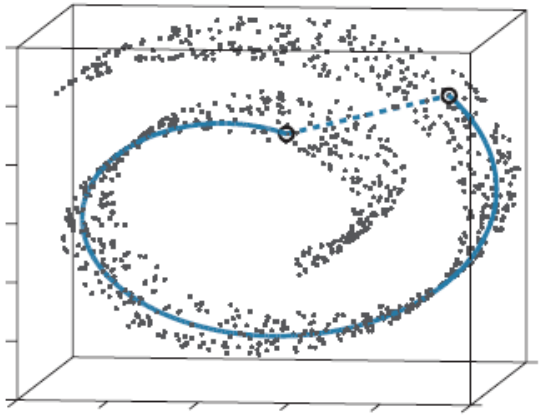


НЕЛИНЕЙНЫЕ МЕТОДЫ ПОНИЖЕНИЯ РАЗМЕРНОСТИ: БАЗОВЫЕ ОПРЕДЕЛЕНИЯ

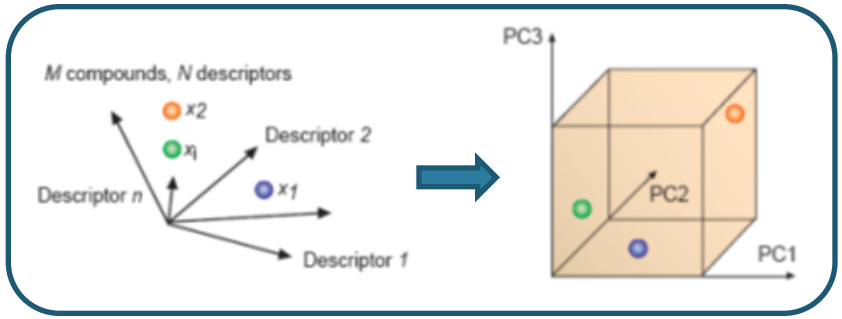


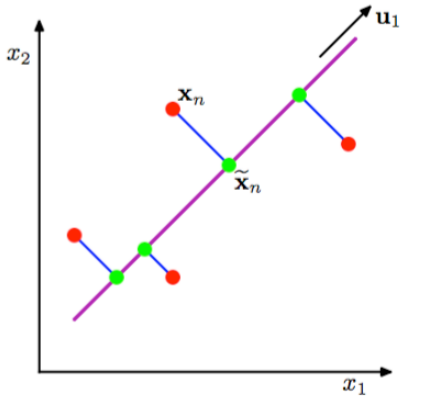


- ***Многообразие (Manifold) -*** топологическое пространство, которое локально выглядит как «обычное» Евклидово пространство**.** В общем случае, любой объект, являющийся условно «плоским» на малых масштабах может рассматриваться как многообразие.
- ***Геодезическое расстояние (Geodesic distance) –*** длина кратчайшей кривой между двумя точками, расположенными на поверхности многообразия (число ребер в кратчайшем пути)



МЕТОД ГЛАВНЫХ КОМПОНЕНТ (PRINCIPAL COMPONENT ANALYSIS)





$$\sum_{i=1}^{\dots} \text{dist}^2(x_i, L_k) \to \min,$$

1. Центрирование данных

$$x_j^{(i)} = x_j - \mu_j, \quad \text{где } \mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)}$$

2. Расчет матрицы ковариаций

$$\Sigma = \frac{1}{m}\sum_{i=1}^{n}\left(x^{(i)}\right)\left(x^{(i)}\right)^T$$

3. Поиск собственных векторов матрицы ковариаций
4. Выбор собственных векторов с наибольшими собственными значениями
5. Проецирование данных на выбранные собственные вектора

$$PC_1 = c_{1,1}x_1 + c_{1,2}x_2 + \cdots c_{1,p}x_p$$
$$PC_2 = c_{2,1}x_1 + c_{2,2}x_2 + \cdots c_{2,p}x_p$$
$$PC_i = c_{i,1}x_1 + c_{i,2}x_2 + \cdots c_{i,p}x_p = \sum_{j=1}^{p} c_{i,j}x_j$$

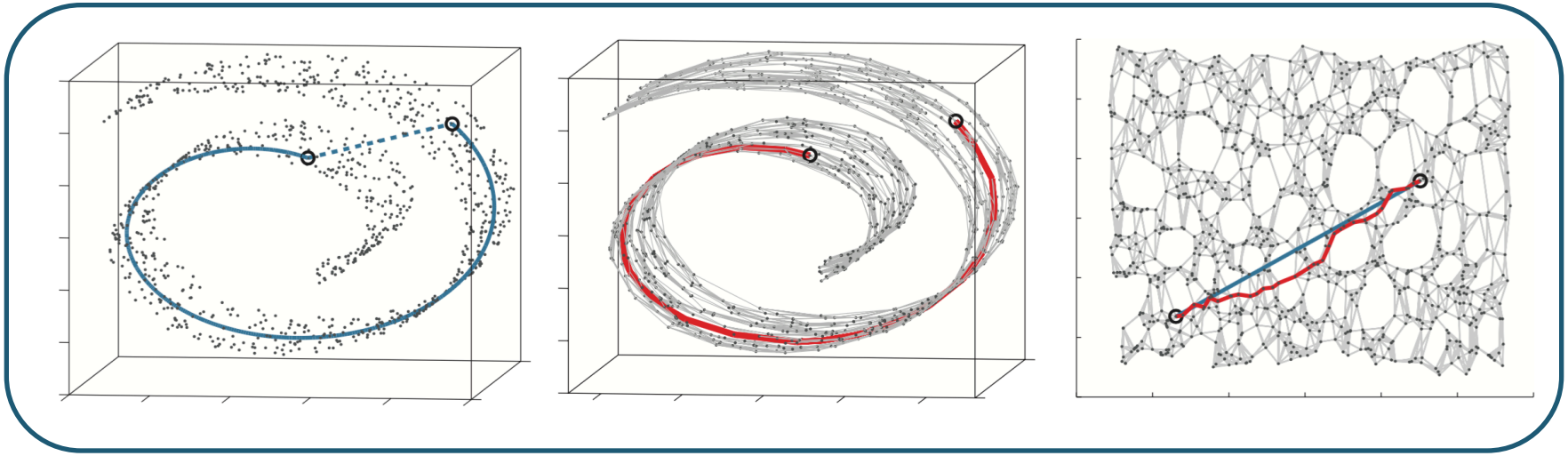
Ограничения метода:

- Линейный (может быть неэффективен при работе с нелинейными данными или данными сложной топологии)
- Сложность интерпретации модели

ISOMAP

Isomap – сочетание алгоритма Флойда — Уоршелла для нахождения кратчайших расстояний между всеми вершинами взвешенного ориентированного графа с классическим многомерным шкалированием.

- Оценка геодезических расстояний между удаленными друг от друга точками.
- Для соседних точек Евклидово расстояние является хорошей аппроксимацией геодезического.
- Для удаленных точек определение расстояния как серии расстояний между близлежащими точками.



J Tenenbaum et al A global geometric framework for nonlinear dimensionality reduction (2000) Science 290 (5500), 2319-2323

ISOMAP

- Определение соседних точек (через расстояния $d_X(i,j)$)
  - Все точки в фиксированном радиусе.
  - K ближайших соседей
- Построить граф окружения G.
  - Каждая точка связана с другой, если та входит в число ее K ближайших соседей.
  - Длины ребер соответствуют Евклидову расстоянию
- Определение геодезических расстояний $d_M(i,j)$ между всеми парами точек путем расчета кратчайших путей между двумя вершинами

  $$\min\{d_G(i,j), d_G(i,k)\ d_G(k,j)\}$$
- Построить пространство меньшей размерности, применив метод многомерного шкалирования MDS к полученной матрице расстояний $D_G = \{d_G(i,j)\}$)

Достоинства:

- Работает с нелинейными данными
- Сохраняет структуру данных

Ограничения:

- Нестабильный (выбор окружения при построении графа)
- Ресурсоемкий

МЕТОД МНОГОМЕРНОГО ШКАЛИРОВАНИЯ (MDS)

Метод анализа данных, позволяющий встраивать точки, соответствующие изучаемым объектам в исходном пространстве в некоторое пространство меньшей размерности так, чтобы попарные расстояния между точками в «новом» пространстве как можно меньше отличались от попарных мер "близости" в исходном пространстве данных.

$$X^\ell = \{x_1, \ldots, x_\ell\} \subset X$$

Мера сходства объектов в исходном пространстве данных: $R_{ij} = \rho(x_i, x_j)$

Евклидовы расстояния в новом пространстве должны как можно более точно соответствовать сходству объектов в исходном пространстве:

$$\phi(Y) = \sum_{ij} (\|x_i - x_j\| - \|y_i - y_j\|)^2$$

raw stress function

или

$$\phi(Y) = \frac{1}{\sum_{ij} \|x_i - x_j\|} \sum_{ij} \frac{(\|x_i - x_j\| - \|y_i - y_j\|)^2}{\|x_i - x_j\|}$$

Sammon cost function

Достоинства:

- Работает с нелинейными данными
- Сохраняет структуру данных

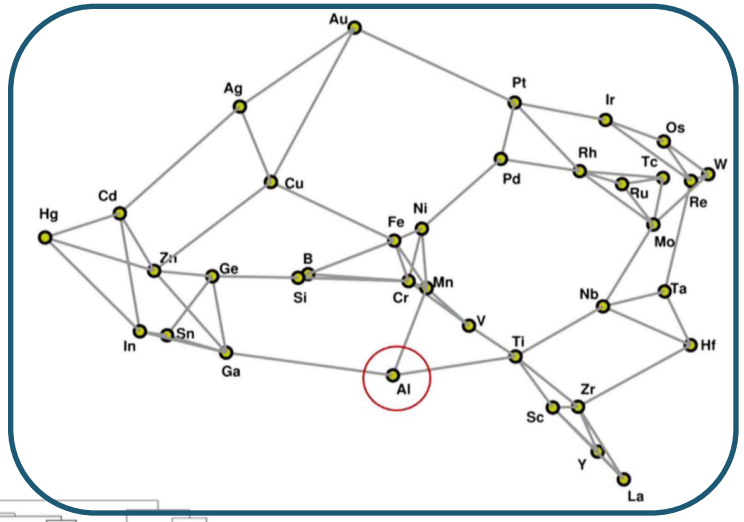
Ограничения:

- Нестабильный
- Ресурсоемкий

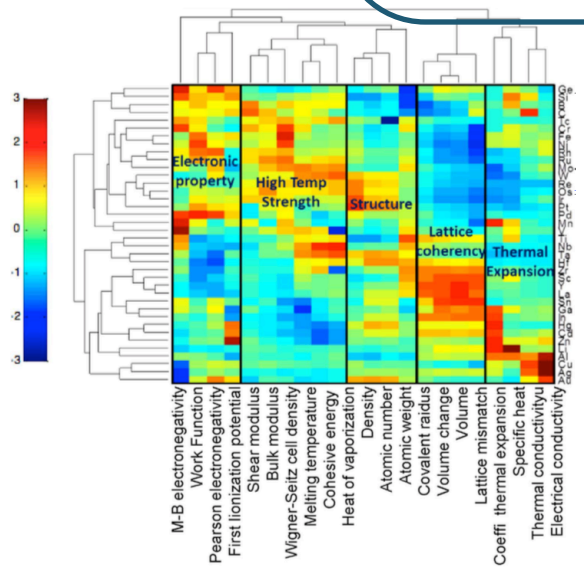
ISOMAP

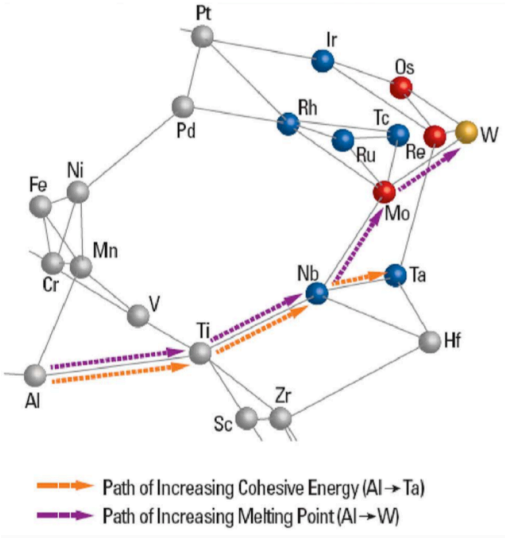
Графический подход для задания меры сходства/различия в разработке жаропрочных сплавов с улучшенными характеристиками

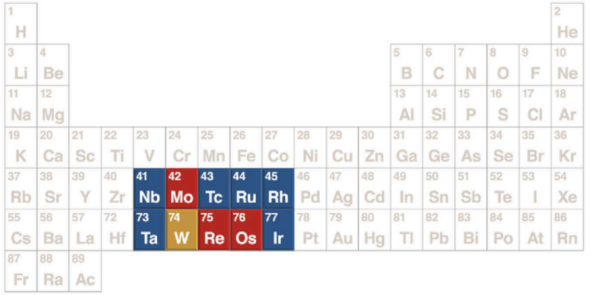
Cobalt-based superalloys $Co_3(Al,X)$



Start	Nearest Neighbor	Cohesive Energy	Nearest Neighbor Map
Al	Al	4.45	
	Ga	4.23	
	Mn	4.07	
	Ti	4.82	
Ti	Al	4.45	
	Nb	5.45	
	Sc	4.61	
	Ti	4.82	
	V	4.77	
	Zr	5.33	
Nb	Hf	5.31	
	Mo	5.05	
	Nb	5.45	
	Ta	5.59	
	Ti	4.82	
Ta	Hf	5.31	
	Nb	5.45	
	Re	5.28	
	Ta	5.59	







Srinivasan, S., S. R. Broderick, et al. (2016). "Mapping Chemical Selection Pathways for Designing Multicomponent Alloys: an informatics framework for materials design." Scientific Reports **5**: 17960.

Stochastic Neighbor Embedding

Сохранение пропорций расстояний при уменьшении размерности

Мера сходства соединений учитывается при помощи преобразования попарных Евклидовых расстояний в условные вероятности

Каждый объект в исходном пространстве описывается нормированными расстояниями (вероятность близости точек *i* и *j* при гауссовом распределении данных с заданным среднеквадратичным отклонением σ) σ выбирается так, чтобы точки в областях с большей плотностью имели меньшую дисперсию:

$$p_{ij} = \frac{\exp(-d_{ij}^2)}{\sum_{k \neq i} \exp(-d_{ik}^2)} \qquad d_{ij}^2 = \frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma_i^2}$$

Сглаженная оценка эффективного числа «соседей»:

Perplexity $Perp(P_i) = 2^{H(P_i)}$

Shannon Entropy $H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}$

Если точки отображения y_i и y_j корректно моделируют сходство между исходными точками высокой размерности x_i и x_j, то и соответствующие условные вероятности будут эквивалентны.

$$p_{ij} = \frac{\exp(-d_{ij}^2)}{\sum_{k \neq i} \exp(-d_{ik}^2)} \qquad q_{ij} = \frac{\exp(-||\mathbf{y}_i - \mathbf{y}_j||^2)}{\sum_{k \neq i} \exp(-||\mathbf{y}_i - \mathbf{y}_k||^2)}.$$

Критерий оценки сходства – дивергенция Кульбека-Лейблера (минимизируем разницу распределения расстояний). SNE минимизирует сумму расстояний для всех точек отображения при помощи метода градиентного спуска.

$$C = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} = \sum_i KL(P_i||Q_i) \qquad \frac{\delta C}{\delta y_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$$

G.E. Hinton and S.T. Roweis. Stochastic Neighbor Embedding. In *Advances in Neural Information Processing Systems*, volume 15, pages 833–840, Cambridge, MA, USA, 2002. The MIT Press.

t-Distributed Stochastic Neighbor Embedding (t-SNE) и Barnes-Hut SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE)

$$q_{ij} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq l} \exp\left(-\|y_k - y_l\|^2\right)}$$

SNE

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}$$

t-SNE

$$\frac{\delta C}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)\left(1 + \|y_i - y_j\|^2\right)^{-1}$$

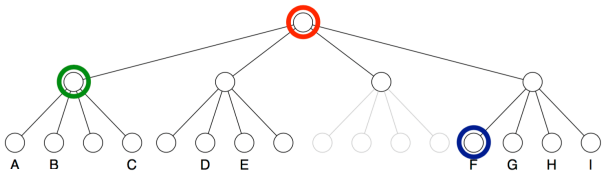
L. Van der Maaten Visualizing Data using t-SNE JMLR 9 (2008) 2579-2605

Barnes-Hut SNE

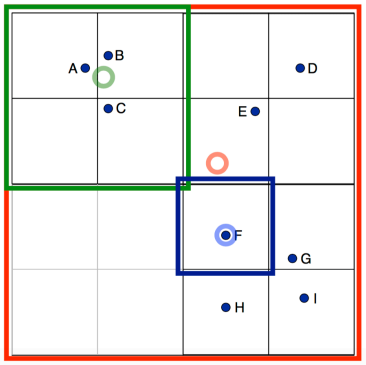
$$\frac{\partial C}{\partial \mathbf{y}_i} = 4(F_{attr} + F_{rep}) = 4\left(\sum_{j \neq i} p_{ij} q_{ij} Z(\mathbf{y}_i - \mathbf{y}_j) - \sum_{j \neq i} q_{ij}^2 Z(\mathbf{y}_i - \mathbf{y}_j)\right)$$

$$Z = \sum_{k \neq l} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}$$

$$F_{rep}$$



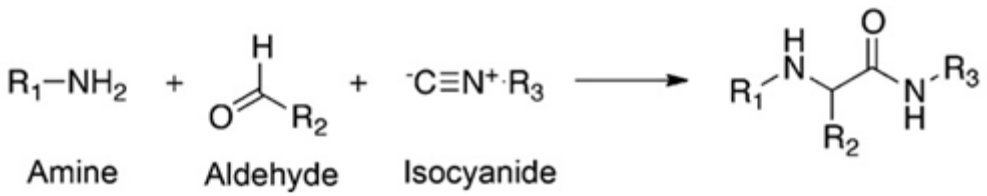
$$\frac{r_{cell}}{\|\mathbf{y}_i - \mathbf{y}_{cell}\|^2} < \theta$$



L. Van der Maaten Accelerating t-SNE using Tree-Based Algorithms JMLR, 15 (2014) 1-21

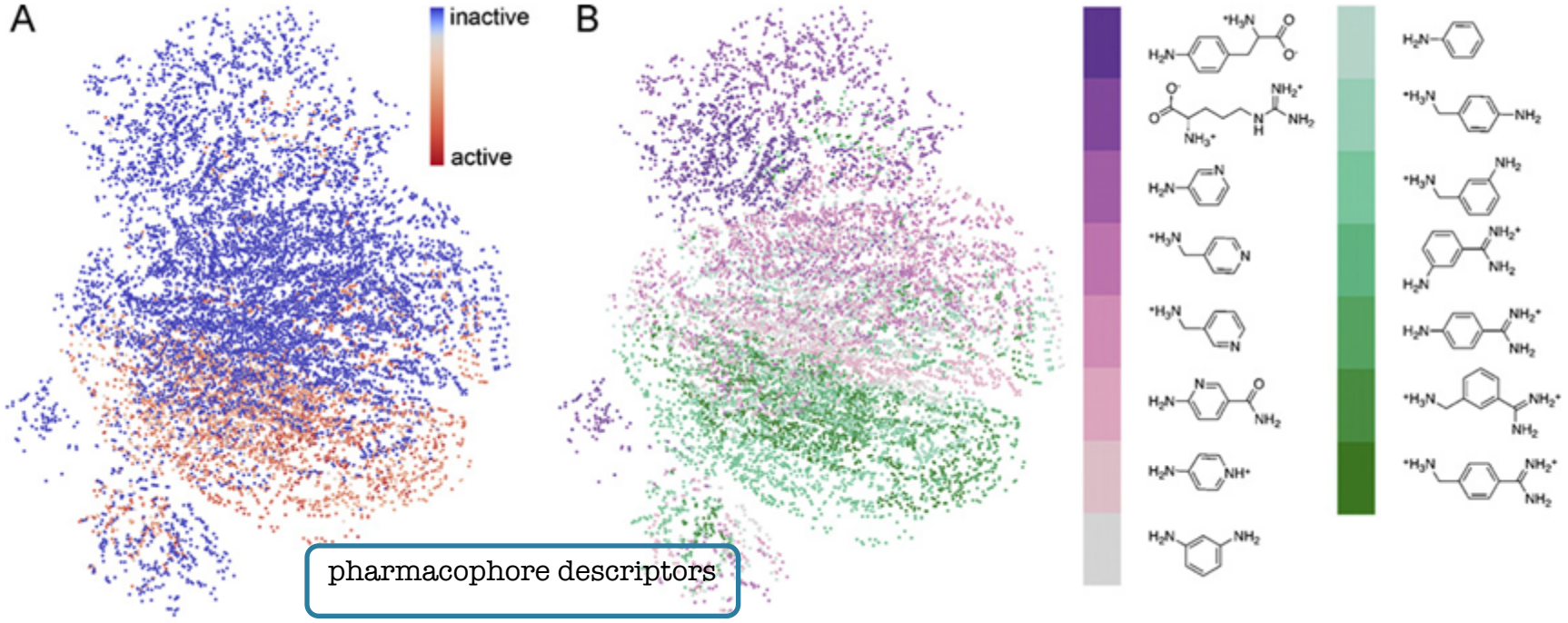
Stochastic Neighbor Embedding

15,840 продуктов (-aminoacyl amide derivatives) трехкомпонентной конденсации Уги (взаимодействие аминов, изоцианидов, карбонильных соединений, органических/неорганических кислот).



Распределение данных по значению биологической активности (ингибирование триптазы)

Распределение данных по подструктуре

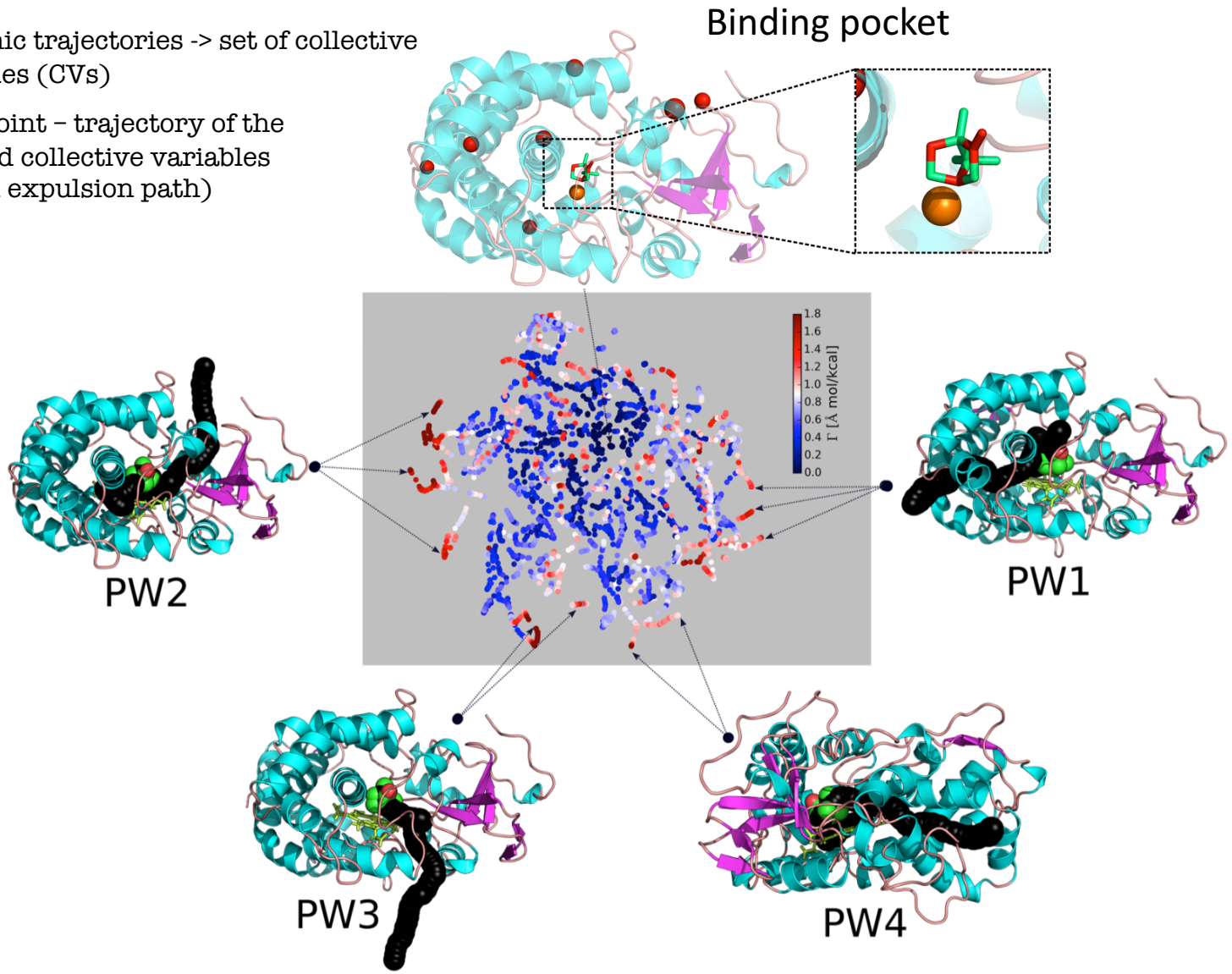


M. Reutlinger, G. Schneider / Journal of Molecular Graphics and Modelling 34 (2012) 108–117

t-SNE as a Tool for Examination Of The Configuration Space Of Cytochrome P450cam Involved In Expulsing Camphor

N-atomic trajectories -> set of collective variables (CVs)

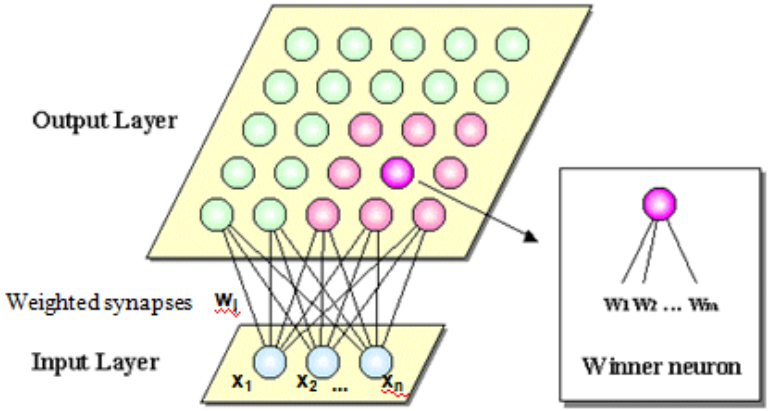
Each point – trajectory of the selected collective variables (ligand expulsion path)

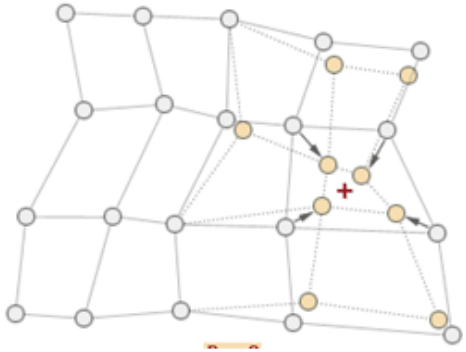


M Rydzewski et al Machine Learning Based Dimensionality Reduction Facilitates Ligand Diffusion Paths Assessment: A Case of Cytochrome P450cam. J. Chem. Theory Comput. 2016, 12, 2110–2120

Метод самоорганизующихся карт Кохонена

Teuvo Kohonen





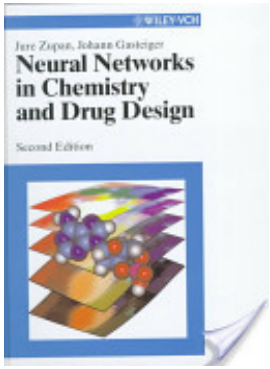
- Нелинейный метод обучения без учителя
- Простота интерпретации моделей
- Сохранение топологии (топология сети в форме решетки, состоящей из нейронов, определяющих дискретное выходное пространство)

•J. Zupan, J. Gasteiger, Neural Networks in Chemistry and Drug Design: An Introduction, Wiley-VCH, Weinheim, 1999.

• S. Anzali, J. Gasteiger, et al. *Perspectives in Drug Discovery and Design,* 1998, 9–11, 273–299

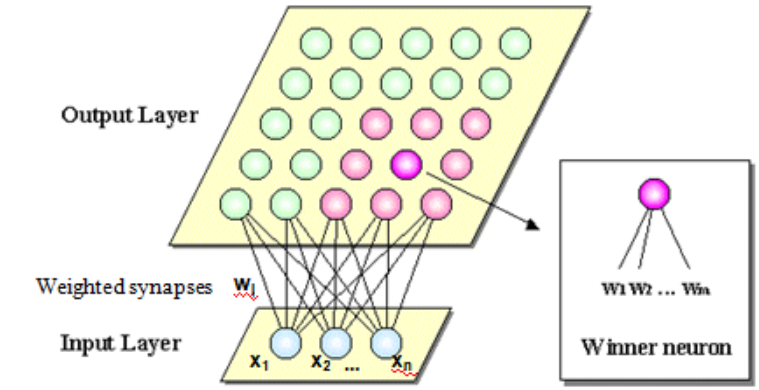
• P. Schneider, Y. Tanrikulu, G. Schneider, Curr. Med. Chem. 2009,16, 258–266.

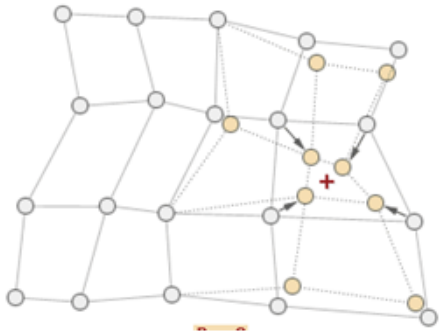
• D. Digles, G. F. Ecker, *Mol. Inf.* 2011, 30, 838 – 846

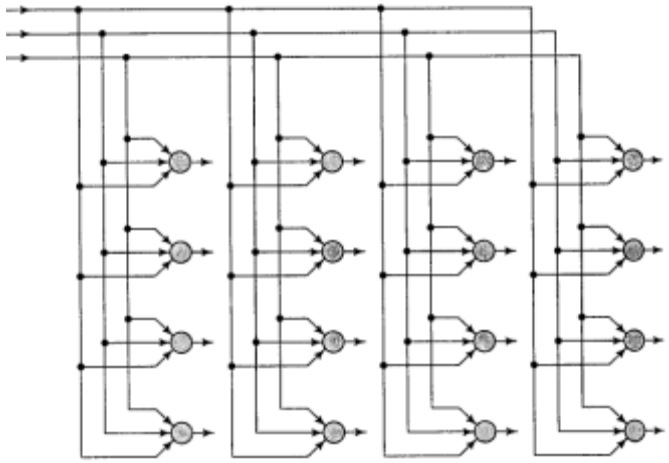


Метод самоорганизующихся карт Кохонена

однослойная нейронная сеть с обучением без учителя, состоящая из компонент, называемых узлами или нейронами.

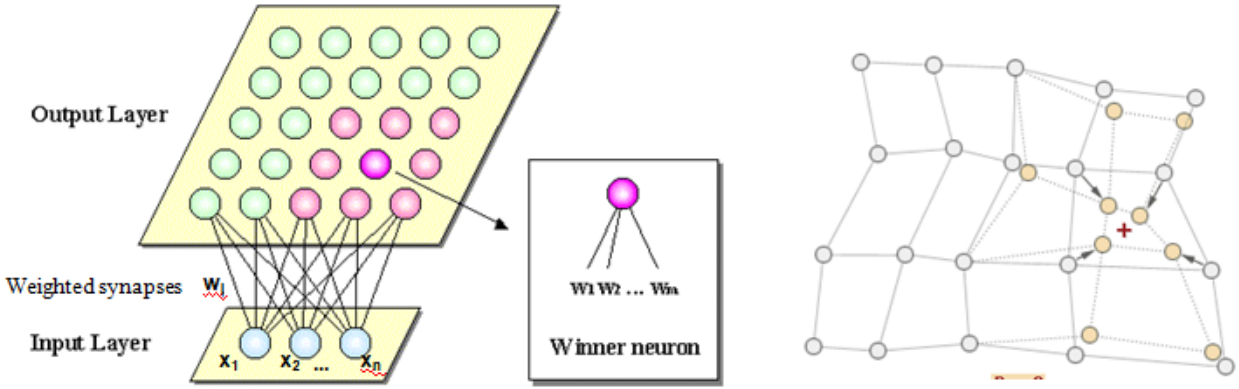






Искусственный нейрон (Математический нейрон Маккалока — Питтса) — узел искусственной нейронной сети, являющийся упрощённой моделью естественного нейрона. Математически, искусственный нейрон обычно представляют как некоторую нелинейную функцию от единственного аргумента — линейной комбинации всех входных сигналов.

Метод самоорганизующихся карт Кохонена



- В процессе обучения координаты узлов (нейронов) приближаются к входным данным
- Для каждого соединения, описанного вектором дескрипторов, выбирается наиболее похожий по вектору веса узел (победивший узел определяет пространственное положение топологической окрестности нейронов). Сходные соединения проецируются в соседние области карты
- Проводится корректировка синаптических весов как для нейрона-победителя, так и для соседних с ним

Самоорганизующиеся карты: алгоритм

Step 1. Инициализация весов.

Step 2. Каждый нейрон в сети получает копию входного вектора

Step 3 (Конкуренция). Поиск нейрона- «победителя», который имеет наименьшее расстояние до входного вектора.

$$d_j = \sqrt{\sum_{i=1}^{M} (x_i - w_{ij})^2}$$

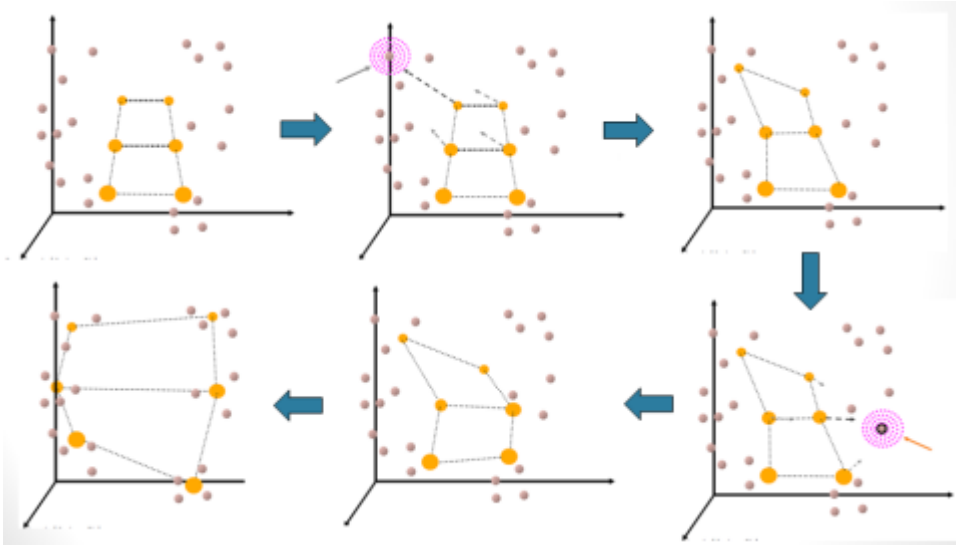
Step 4 (Кооперация, синаптическая адаптация). Для победившего нейрона и его окружения модифицируются веса:

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t) h_{ij}(t) \cdot (x_i - w_{ij}(t)),$$

где ***h(t)*** – функция топологической окрестности, $\eta(n)$ - параметр скорости обучения

Step 5. Повторить **Steps 2-4** для всех входных векторов

Step 6. Повторить **Step 5** заданное количество раз



Самоорганизующиеся карты Кохонена: алгоритм

Процесс кооперации

Определение топологической окрестности

Топологическая окрестность является симметричной относительно точки максимума
Амплитуда топологической окрестности монотонно уменьшается с увеличением расстояния (типичным примером является функция Гаусса)

Уменьшение размера топологической окрестности со временем

$$h_{j,i(\mathbf{x})}(n) = \exp\left(-\frac{d_{j,i}^2}{2\sigma^2(n)}\right), \quad n = 0, 1, 2 \ldots,$$

функция топологической окрестности

$$\sigma(n) = \sigma_0 \exp\left(-\frac{n}{\tau_1}\right), \quad n = 0, 1, 2, \ldots,$$

ширина функции топологической окрестности

Процесс адаптации

Выбор параметра скорости обучения

$$\eta(n) = \eta_0 \exp\left(-\frac{n}{\tau_2}\right), \quad n = 0, 1, 2, \ldots,$$

Самоорганизующиеся карты Кохонена: этапы адаптивного процесса

Этап самоорганизации (упорядочения)

- Топологическое упорядочение векторов весов (до 1000 итераций)
- Выбор параметра скорости обучения (η инициализировать значением 0.1, временной параметр τ_2 1000) и функции окрестности (должна изначально охватывать практически все нейроны сети) являются определяющими

Этап сходимости

- Требуется для точной подстройки карты признаков
- Количество итераций как правило в 500 раз превышает количество нейронов сети
- Функция окрестности должна охватывать только ближайших соседей
- Параметр скорости обучения должен быть инициализирован достаточно малым значением

Самоорганизующиеся карты Кохонена (SOM): ограничения и сильные стороны

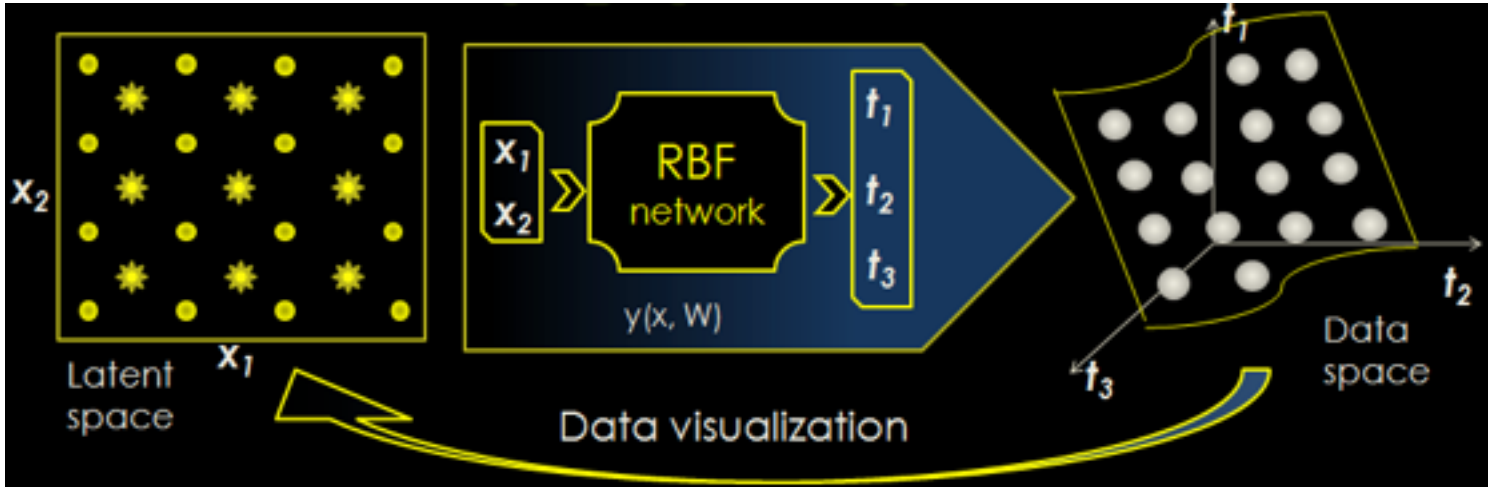
Сильные стороны:

- Возможность работы с многомерными данными сложной топологии, поиск нелинейных взаимосвязей
- Сохраняет топологию многомерного пространства данных

Ограничения:

- Отсутствие оптимизируемой функции -> отсутствие гарантии сходимости
- Отсутствие теоретического обоснования для выбора ряда внутренних параметров метода
- Зависимость от инициализации

Генеративные Топографические Карты (Generative Topographic Maps)

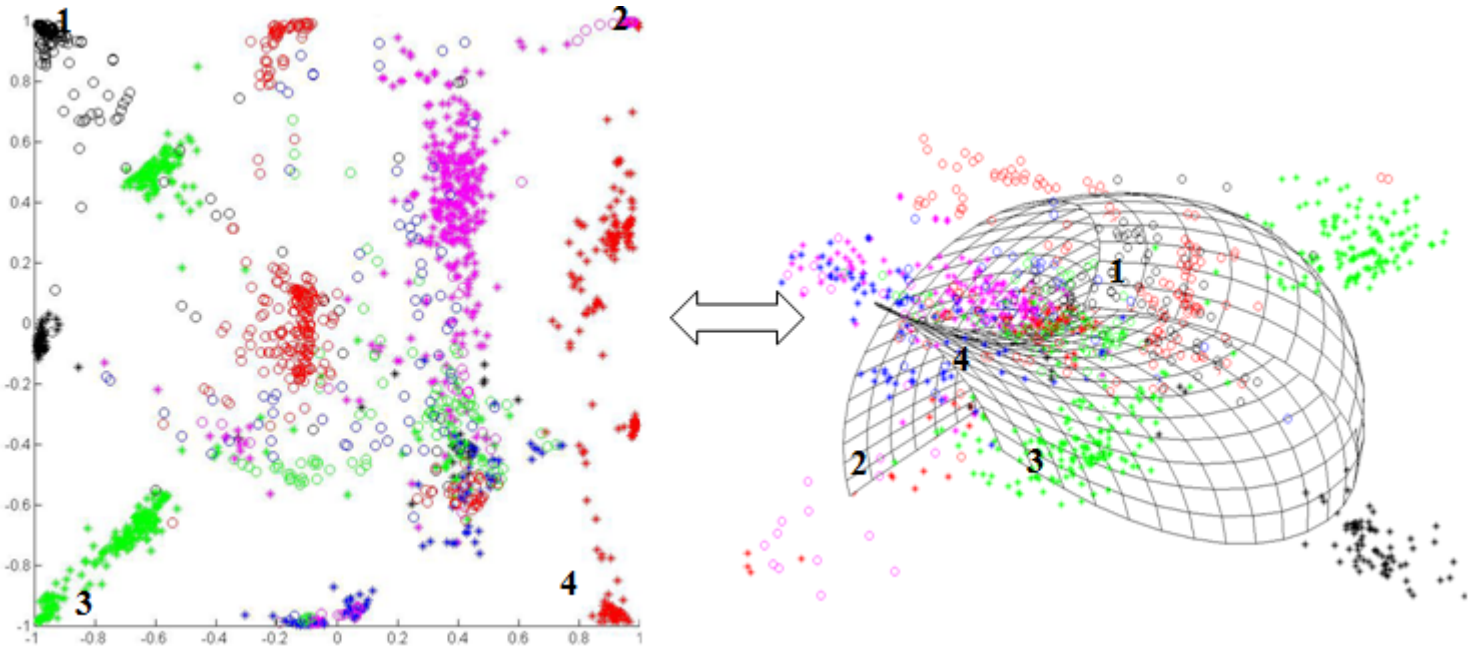


- Метод генеративных топографических карт может рассматриваться как вероятностная модификация метода самоорганизующихся карт
- GTM генерирует распределение вероятности в пространстве данных посредством смеси Гауссиан, встроенных в него и аппроксимирующих плотность данных
- Обучающий алгоритм максимизирует функцию правдоподобия, сходимость контролируется EM-алгоритмом.

$$\mathcal{L}(\mathbf{W}, \beta) = \sum_{n=1}^{N} \ln \left\{ \frac{1}{K} \sum_{i=1}^{K} p(\mathbf{t}_n | \mathbf{x}_i, \mathbf{W}, \beta) \right\}$$

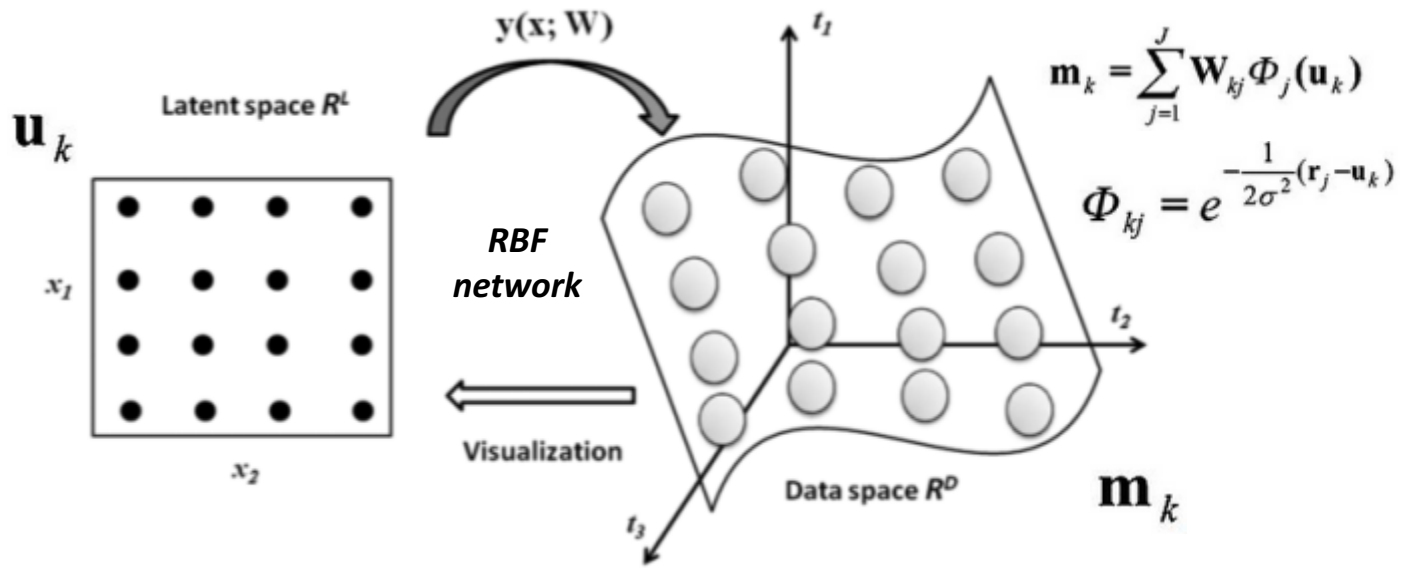
- C. M. Bishop *Pattern Recognition and Machine Learning*, 2006 Springer
- D. M. Maniyar, I. T. Nabney, et al. *J. Chem. Inf. Model.,* 2006, 46, 1806-1818

Генеративные Топографические Карты (Generative Topographic Maps)

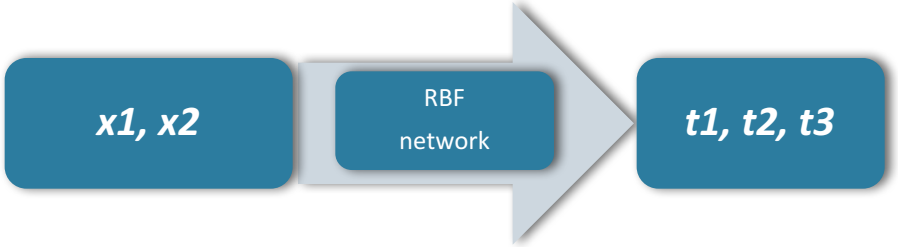


- GTM соотносит латентное пространство с 2D многообразием (манифолдом), встроенным в многомерное пространство данных.
- Визуализация осуществляется проецированием точек данных на многообразие.

Генеративные Топографические Карты (Generative Topographic Maps)



$$\mathcal{L}(\mathbf{W}, \beta) = \sum_{n=1}^{N} \ln \left\{ \frac{1}{K} \sum_{i=1}^{K} p(\mathbf{t}_n | \mathbf{x}_i, \mathbf{W}, \beta) \right\}$$



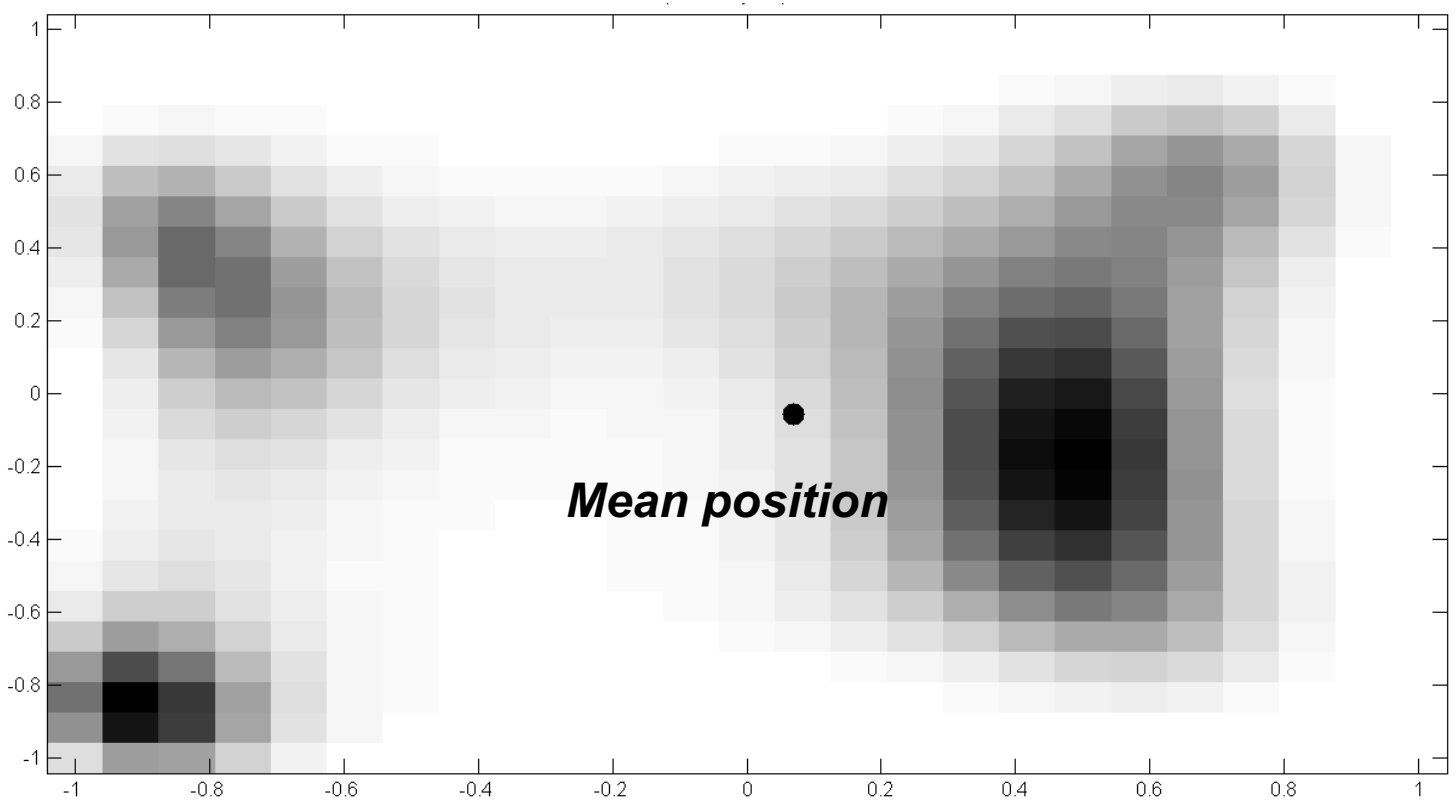
EM алгоритм

- E-step:
  - responsibility of latent point x_k for data point t_n

$$r_{kn} = p(\mathbf{x}_k | \mathbf{t}_n, \mathbf{W}, \beta) = \frac{p(\mathbf{t}_n | \mathbf{x}_k, \mathbf{W}, \beta) p(\mathbf{x}_k)}{\sum_{k'} p(\mathbf{t}_n | \mathbf{x}_{k'}, \mathbf{W}, \beta) p(\mathbf{x}_{k'})}$$

  - $p(x_k)$ constant (1/K)
- M-step:
  - r_{kn} used as weights to update β and W
  - “move each component of the mixture towards data points for which it is most responsible”

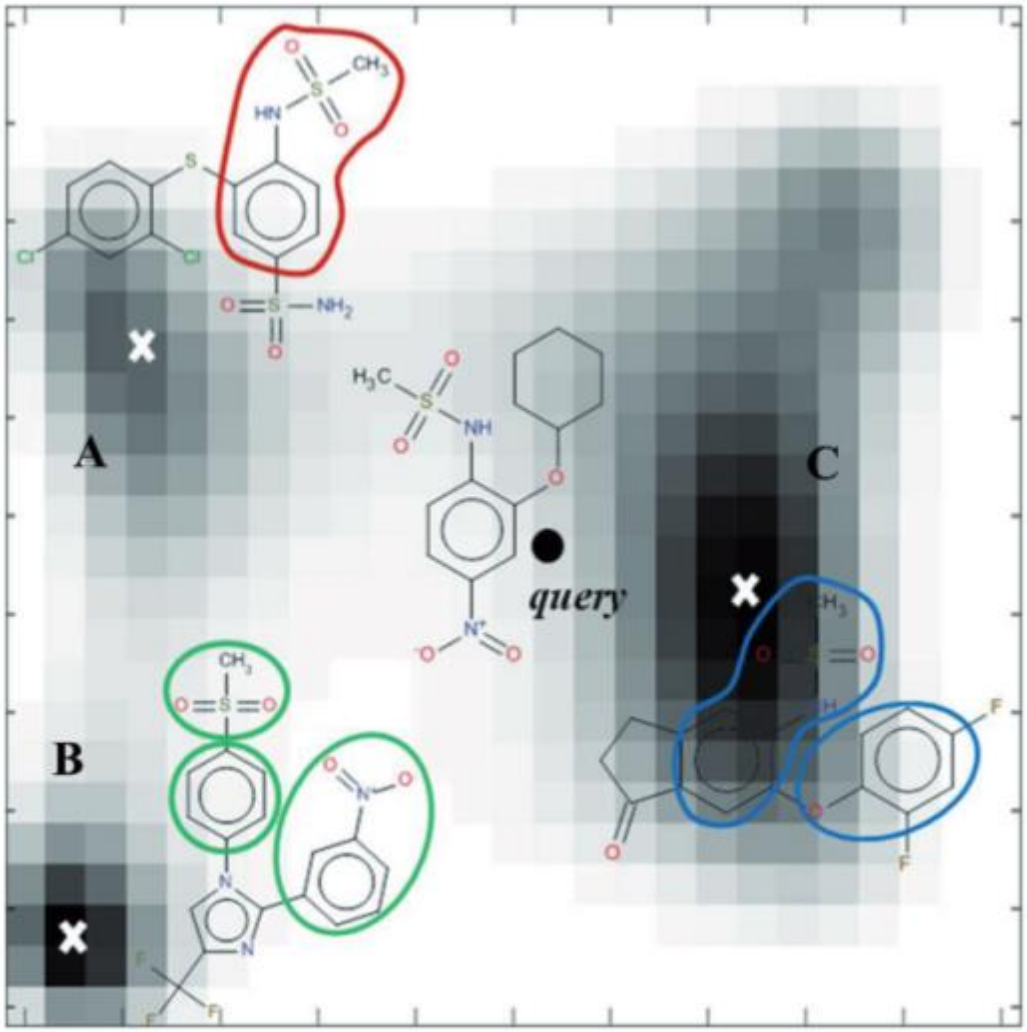
Генеративные Топографические Карты: мультимодальное распределение вероятности



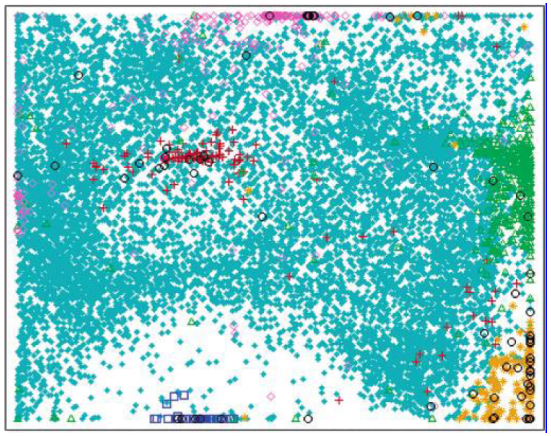
Проекция объекта на карту определяется распределением вероятности между латентными точками:

$$\mathbf{x}_n^{\text{mean}} = \sum_k^K \mathbf{x}_k p(\mathbf{x}_k | \mathbf{t}_n)$$

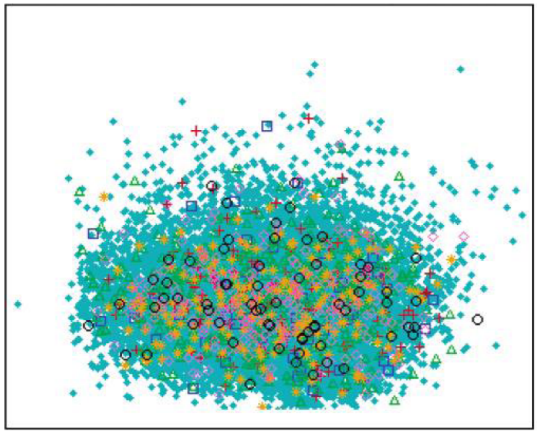
Генеративные Топографические Карты: мультимодальное распределение вероятности



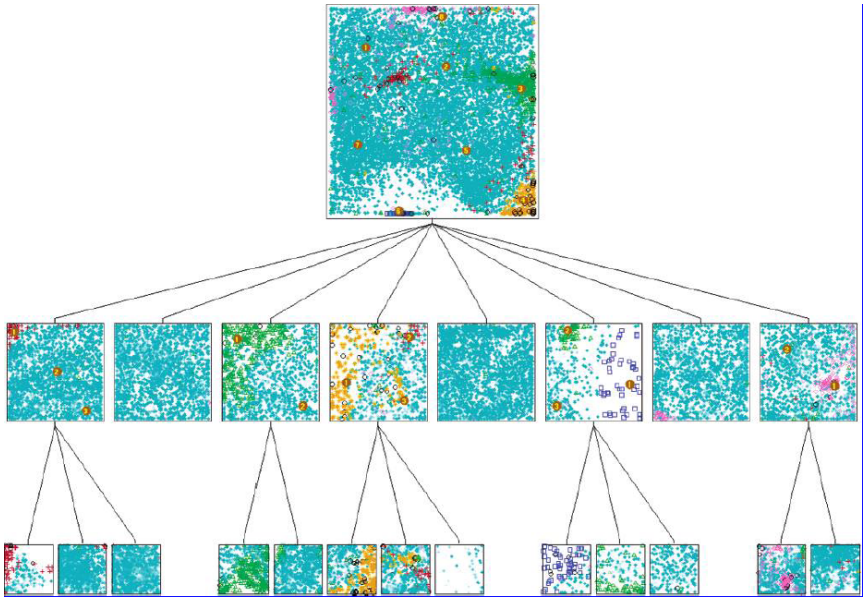
Метод Генеративных Топографических Карт: примеры использования



GTM



SOM



Label Description	Marker	Compounds
Not active in any screen	•	10769
Active for peptidergic type1	+	118
Active for peptidergic type2	*	181
Active for aminergic type1	□	50
Active for aminergic type2	△	409
Active for kinase	◇	206
Active for more than 1 screen	○	66

The data set provided by Pfizer is composed of the data for five different targets (11 800 compounds were randomly selected from 1 M)

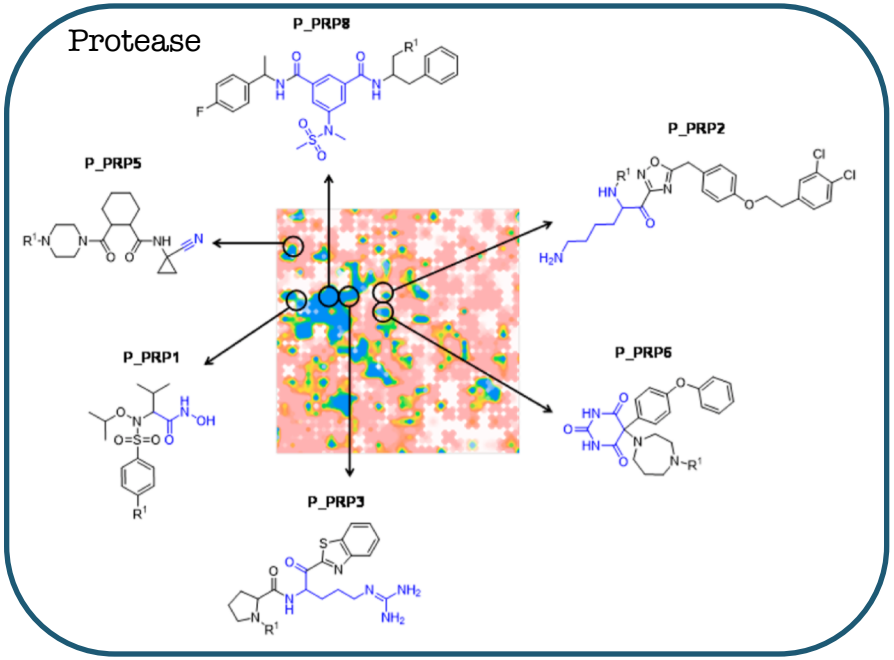
D. M. Maniyar et al J. Chem. Inf. Mod. 2006, 46, 1806-1818

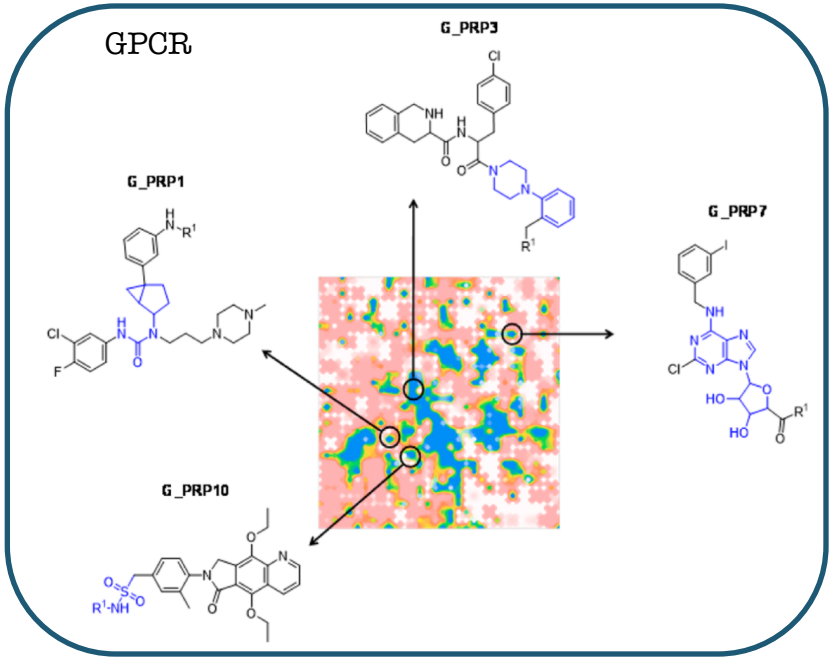
Метод Генеративных Топографических Карт: идентификация привилегированных подструктур

Descriptors: different types of pharmacophore-colored descriptors (ISIDA)

Data: subsets of compounds with Ki and/or IC50 values against human targets were assembled from ChEMBL.

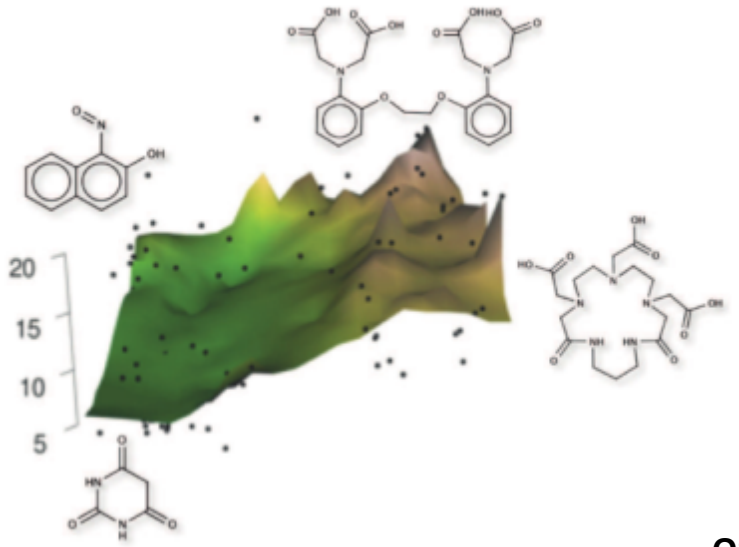
superfamily	target family	#CPDs
proteases	serine proteases	7585
	metallo proteases	4131
	cysteine proteases	3227
	aspartic proteases	3068
	threonine proteases	165
kinases	serine threonine kinases	10,804
	tyrosine kinases	9907
	PI3/PI4 kinases	1982
GPCRs	short peptide receptors	14,472
	monoamine receptors	14,101
	lipid-like ligand receptors	7613
	nucleotide-like receptors	5811
	chemokine receptors	5042





S Kayastha et al Privileged Structural Motif Detection and Analysis Using Generative Topographic Maps *JCIM* 2017 *57* (5), 1218-1232

Генеративные топографические карты: QSAR модели



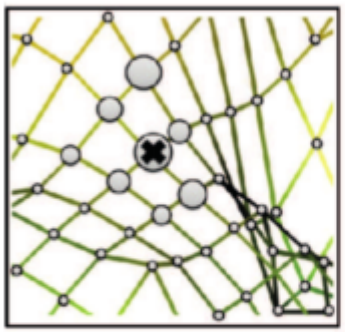
$$\overline{A}_k = \frac{\sum_{n=1}^{N} A_n R_{kn}}{\sum_{n=1}^{N} R_{kn}}$$

QSAR модели на основе глобального ландшафта активности:

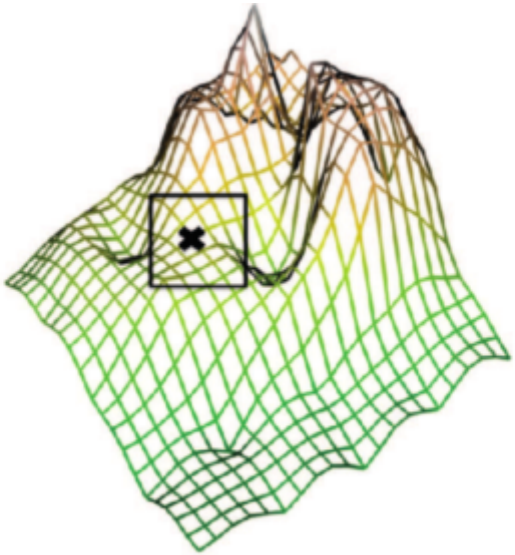
$$\widehat{A}_j = \sum_k \overline{A}_k R_{kj}$$

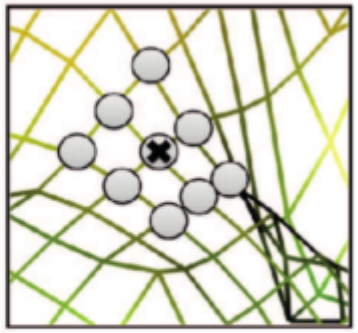
QSAR модели на основе локального окружения:

$$\widehat{A}_j = \frac{\sum_v \overline{A}_v}{V}$$



(a)





(b)

Mol. Inf. 2015,34, 348–356

Метод Стохастического Триплетного Встраивания (Stochastic Triplet Embedding)

В основе метода лежит принцип, сходный с системой человеческих суждений при относительной оценке, представляемых для анализа объектов:
“Предмет А больше похож на В или на С?”

Функция сходства $s(\mathbf{z}_i, \mathbf{z}_j)$ заменяется набором триплетов индексов:

$$\mathcal{T} = \{(i, j, \ell) | \ \mathbf{z}_i \text{ is more similar to } \mathbf{z}_j \text{ than } \mathbf{z}_\ell\} \qquad s(\mathbf{z}_i, \mathbf{z}_j) < s(\mathbf{z}_i, \mathbf{z}_\ell)$$

Задача сводится к максимизации вероятности (верности утверждения) по всем тройкам соединений в обучающем наборе данных:

$$\max_X \sum_{\forall (i,j,l) \subseteq \tau} \log p_{ijl};$$

Где вероятность соответствия триплета стохастическому правилу отбора:

$$p_{ijl} = \frac{\left(1 + \frac{x_i - x_j^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}}{\left(1 + \frac{x_i - x_j^2}{\alpha}\right)^{-\frac{\alpha+1}{2}} + \left(1 + \frac{x_i - x_l^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}};$$

Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing, 2012

Вопросы?