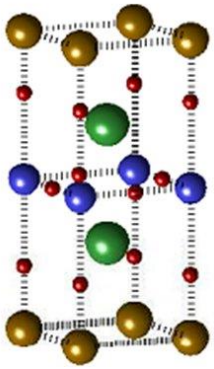
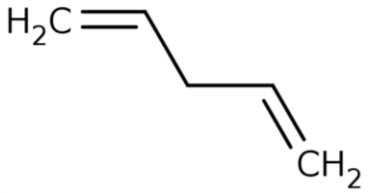
Топологическая теория анализа хемографов как перспективный подход к имитационному моделированию квантово-механических свойств молекул

Торшин Иван Юрьевич

Рудаков Константин Владимирович





…
Tc
LUMO
ZPVE
U
…

«структура» → «свойство»

ММРО, 2021

1. Комбинаторная теория разрешимости
2. «Топологический» анализ данных
3. Метрический анализ данных (в т.ч. порождение проблемно-ориентированных метрик, задачи согласования значений метрик)

Содержание

- Имитационное моделирование и квантовая механика
- Основы **теории анализа размеченных графов**
- Основы **топологической теории анализа данных**
- Проблемно-ориентированная теория для оценочных вычислений квантово-механических свойств молекул по структурной формуле.
- **Интерпретации** в рамках КМ
- **Апробация алгоритмов** на выборке из 134000 молекул
- **Интерпретации** в терминах теории химической связи

"Химия - в существенной степени корреляционная наука, базирующаяся на установлении соответствий структура-свойство для последующего построения алгоритмов прогнозирования"

Степанов Н.Ф. Квантовая механика, 2001

- Критерии применения методов имитационного моделирования (ИМ) сложных систем
  1. не разработаны аналитические модели,
  2. не разработаны точные методы решения аналитических моделей
  3. имеющиеся решения неустойчивы
  4. вычисления неприемлемо длительны
- Парадигма "машинного обучения" - полезный инструмент для ИМ

Оценки качества ИМ

a) точность моделей,
b) обобщающая способность,
c) скорость вычислений,
d) интерпретируемость получаемых результатов.

чем ниже точность/обобщающая способность, тем более результаты моделирования оторваны от реальности

принципиально важна для анализа/скрининга больших выборок молекул, молекул большого размера

в терминах теории химической связи и структурной химии, на доступном для химиков-практиков уровне

Квантовая механика (КМ)

- Ярчайший пример успешного применения теории вероятностей, теории операторов, теории групп и функционального анализа в теоретической физике.
- Тем не менее
  - точное аналитическое решение уравнения Шрёдингера имеется только для атома водорода (п. 2 выше).
  - КМ вычисления затратны даже для систем малого размера (десятки-сотни атомов, п. 4. выше).
- Поэтому, перспективен поиск моделей ИМ оценочного для расчёта КМ-свойств молекул.

Аксиоматика в основе математических конструкций КМ

- **Постулат 1**. *Состояние квантовой системы из N микрочастиц (электронов и ядер) полностью определяется волновой функцией от радиус-векторов частиц и времени.* $\psi \in \mathfrak{K}$ $\overline{A} = \langle \psi | \hat{A} | \psi \rangle$ $\psi(\mathbf{x}, t) : R^{3N+1} \rightarrow R$
- **Постулат 2**. *Каждая наблюдаемая физическая величина A представима в виде линейного оператора* $\hat{A}\psi$*, применяемого к ψ-функции системы для получения значений величины A.* Среднее значение величины A - $\overline{A} = \int_{R^{3N}} \psi^* \hat{A} \psi d\mathbf{x}$
- **Постулат 3**. *Изменение волновой функции во времени определяется уравнением Шредингера (УШ). В стационарном виде УШ представляет собой уравнение на собственные значения гамильтониана.*
- **Постулат 4**. *Электроны в квантовой системе неразличимы.*

$$\hat{H}\psi = i\hbar \frac{\partial \psi}{\partial t} \quad \hat{H} = \hat{T} + V \quad \hat{T} = \sum \hat{T}_i \quad \hat{T}_i = \frac{\hbar^2}{2m_i}\Delta_i \quad \Delta_i = \frac{\partial}{\partial x_i^2} + \frac{\partial}{\partial y_i^2} + \frac{\partial}{\partial z_i^2} \quad V(\mathbf{x}, t) : R^{3N+1} \rightarrow R \quad \vec{r}_i = (x_i, y_i, z_i) \quad \mathbf{R} = (\vec{R}_\alpha)$$

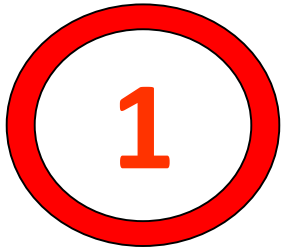
$$\hat{H}\psi(\mathbf{r}, \mathbf{R}) = E\psi(\mathbf{r}, \mathbf{R}) \quad \hat{H}_{\text{адиаб}} = \frac{1}{2}\sum_i \frac{\hbar^2}{2m_e}\Delta_i(\mathbf{r}) + V_{ee}(\mathbf{r}) + V_{en}(\mathbf{r}, \mathbf{R}) + V_{nn}(\mathbf{R}) \quad V_{ee}(\mathbf{r}) = \frac{1}{2}\sum_i \sum_{j \neq i} \frac{1}{d_{ij}} \quad d_{ij} = \left\| \vec{r}_i - \vec{r}_j \right\| \quad V_{en} = -\sum_{i,\alpha} \frac{Z_\alpha}{R_{\alpha i}} \quad R_{\alpha i} = \left\| \vec{R}_\alpha - \vec{r}_i \right\|$$

$$V_{nn} = \sum_{\alpha,\beta} \frac{Z_\alpha Z_\beta}{R_{\alpha\beta}} \quad R_{\alpha\beta} = \left\| \vec{R}_\alpha - \vec{R}_\beta \right\| \quad C = (\mathbf{r}_j) \subset R^3 \quad \hat{h}_i \psi_{1,k}(\vec{r}_i, \mathbf{R}) = e_{ik} \psi_{1,k}(\vec{r}_i, \mathbf{R}) \quad \hat{h}_i = -\frac{\hbar^2}{2m_e}\Delta_i - \sum_\alpha \frac{Z_\alpha}{R_{\alpha i}} + V_i(i) \quad \hat{H}_e = \sum_i \hat{h}_i \quad E_k = \sum_i e_{ik}$$

$$V_i(i) = \frac{1}{2}\sum_{i \neq j} \frac{1}{d_{ij}}$$

$C = (\mathbf{R}) \rightarrow$ **M(C)=(d_{ij}(C))** $\rightarrow$ **Теория анализа хемографов**

Основы теории анализа размеченных графов



Установление изоморфизма графов, полнота инвариантов графа

- Граф - элемент множества $\Gamma = \{(V, E) \mid V \subseteq \mathbf{N}, E \subseteq \mathbf{N}^2\}$
  «множество всех графов»
  N - натуральный ряд.
- Графы G_1, G_2 *тождественны* ($G_1 = G_2$), если $V_1 = V_2$ и $E_1 = E_2$
- Инцидентность вершины *v* и ребра *e*

$$v \bullet e \equiv (e = (v_1, v_2), (v = v_1) \vee (v = v_2))$$

- Графы G_1, G_2 ***изоморфны*** ($G_1 \simeq G_2$) - существует взаимно-однозначное соответствие между их вершинами и рёбрами, сохраняющее смежность вершин и инцидентность рёбер.

Th. $G_1 \simeq G_2 \Leftrightarrow \exists \mu_I : V_1 \to V_2 \mid G_2 = (\{\mu_I(v), v \in V_1\}, \{(\mu_I(u), \mu_I(v)), (u, v) \in E_1\})$

Инварианты графов

- *Инвариант* графа — числовая характеристика графа или упорядоченный список таких характеристик (кортеж), значение которой одинаково для каждого элемента произвольного **класса изоморфных графов**.

  $\iota : \mathbf{\Gamma} \to \mathrm{R}^{\mathrm{n}}, \mathrm{n} \in \mathbf{N}$ $\quad \forall \mathrm{a} \in \mathbf{\Gamma} : \mathrm{b} \in \mathbf{I}(\mathrm{a}) \Rightarrow \iota(\mathrm{b}) = \iota(\mathrm{a})$
- *Элементарный инвариант* $\iota : \mathbf{\Gamma} \to \mathrm{R}$
- *Кортеж-инвариант* $\iota : \mathbf{\Gamma} \to \mathrm{R}^{\mathrm{n}}, \mathrm{n} \geq 2$
- В теории графов - десятки элементарных инвариантов:
  - сумма длин минимальных цепей между каждой парой вершин (индекс Винера)
  - кликовое число
  - число компонент связности графа
  - ….
- ***Условие полноты инварианта:*** $\forall \mathrm{a} \in \mathbf{\Gamma} : \mathrm{b} \in \mathbf{I}(\mathrm{G}) \Leftrightarrow \iota(\mathrm{a}) = \iota(\mathrm{b})$

О «локальной полноте» инвариантов

- Утверждения теории графов – по отношению к бесконечному множеству **Г**
- *Локальные формы утверждений о графах* формулируются по отношению к определенному *конечному* $\mathrm{P} \subset \Gamma$
- ***Целесообразность:*** *возможность комбинаторного тестирования выполнимости утверждений над* $\mathrm{P} \subset \Gamma$

В рамках **комбинаторной теории разрешимости**, графы рассматриваются как объекты, а их кортеж-инварианты - как вектора признаковых описаний объектов.

Основные положения **комбинаторной теории разрешимости/регулярности** задач распознавания/классификации

1

$$I_i \xrightarrow[\mathbf{F}]{} I_f$$

I. Определены *множество* ***признаковых*** *описаний и* ***классы*** *объектов*.

II. *Задано* ***конечное множество объектов*** («множество прецедентов»), описываемое совокупностью матрицы информации (признаки объектов) и информационной матрицы (принадлежность объектов к классам).

III. *Множество объектов* ***непротиворечиво***, т.е. для произвольного набора признаков объекта существует только один ответ в информационной матрице.

IV. Непротиворечивость *множества объектов при* **заданном признаковом описании** *является необходимым и достаточным условием* ***разрешимости задачи*** (т.е. существования отображения из множества матриц информации во множество информационных матриц).

V. Разрешимость задачи при заданном множестве прецедентов может быть достигнута на определенных ***подмножествах*** **множества признаков**.

VI. При заданных подмножестве признаков и множестве прецедентов, ***локальная*** выполнимость критерия разрешимости устанавливается путем **комбинаторного тестирования** на исследуемом множестве прецедентов.

VII. Подмножества признаков могут выбираться на основании **«информативности»**

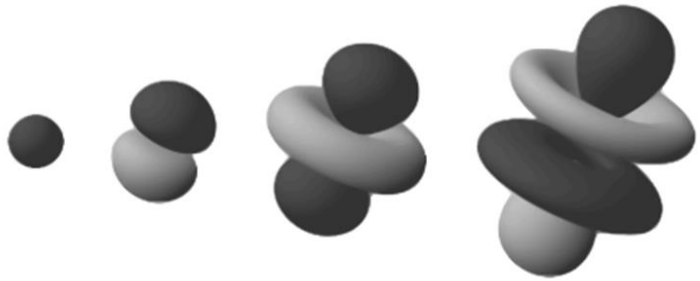
Графы специального вида – хемографы (χ-графы)

- **Определение.** *Хемографом X будем называть **конечный,** связный, неориентированный, **размеченный** граф без петель , с кликовым числом не превышающим 3 и для которого выполнена аксиома кратности связей.*

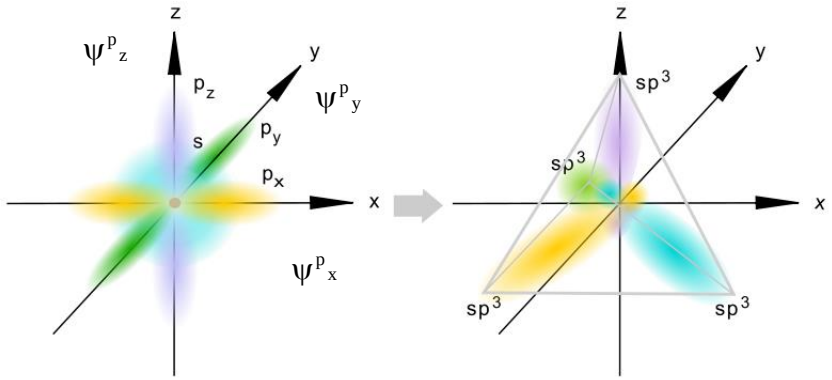
Аксиома кратности химических связей $d+w_{max}=v_A+1.$

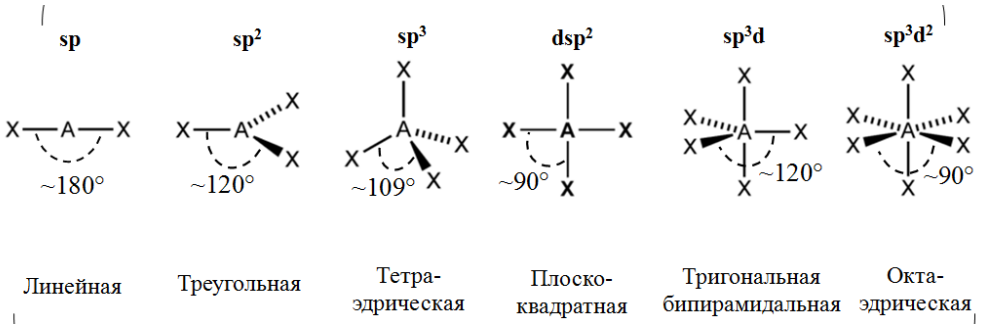
- d - число связей атома
- w_{max} – макс. кратность связи
- v_A - макс. валентность атома

$$-\frac{\hbar^2}{2m}\Delta\psi(\vec{r})+U(\vec{r})\psi(\vec{r})=E\psi(\vec{r})$$



$$c_i(\psi_i^s+\sqrt{3}\psi_i^p)$$





Аксиома кратности в терминах теории графов

$$d+w_{max}=v_A+1.$$

d - «число связей атома»

w_{max} – «макс. кратность связи»

v_A - «макс. валентность»

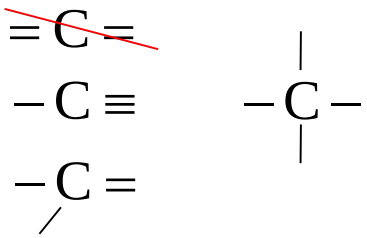
число смежных атому вершин в данном графе

максимальный элемент в строке соответствующей взвешенной матрицы смежности данного графа

максимально возможное число ребер, инцидентных любой вершине в любом графе, соответствующей «типу атома» (т.е. метке вершины)

- Аксиома кратности:
  - Общее число ребер вершины с заданной меткой постоянно, а число инцидентных вершин - может изменяться.

Пример: метка «С»



Ограничения на словарь разметки и подмножества возможных сочетаний меток

Разметки χ-графов

Множество (алфавит) меток Y

Функция разметки вершин $\mu_V : V \rightarrow Y$.

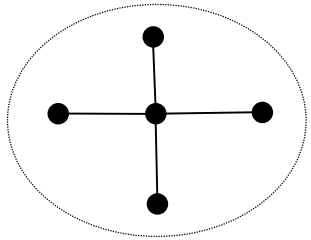
- В качестве меток можно использовать
  - химические типы атомов («C», «N», «O» и т. д.)
  - гибридные состояния атомов
  - максимальную валентность атома
  - заряды атомов
  - *комбинированные метки*
  - …

Аксиома кратности связей

Сочетания меток: χ-цепи

- Для порождения признаковых описаний хемографов вводятся понятия χ-цепей и χ-узлов.
  - X – размеченный граф, *множество меток* $Y=\{\upsilon_1, \upsilon_2, \ldots, \upsilon_{n(Y)}\}$, *функция разметки* $\mu_V : V \to Y$.
  - множество всех перестановок над Y - $\ddot{Y} = \bigcup_{n=1}^{\infty} Y^n$
    $y^1 = (y^1_1, y^1_2, \ldots, y^1_i \ldots, y^1_n)$
    $y^2 = (y^2_1, y^2_2, \ldots, y^2_i \ldots, y^2_n)$ $\quad y^1, y^2 \in \ddot{Y}$
  - ***χ-цепь*** α - элемент *множества χ-цепей,* $\alpha \in \tilde{Y}$
    $\tilde{Y} = \{\{y^1, y^2\}, y^1, y^2 \in \ddot{Y} \,\|\, y^1 \,|=|\, y^2 \,| = n, \forall i = 1..n : y^1_i = y^2_{n-i+1}\}$
  - *Множество всех χ-цепей длины n* $\tilde{Y}^n \subset \tilde{Y}$

Сочетания меток: χ-узлы



- *k-элементными сочетаниями* над Y - элементы множества $\sigma Y^k = \{\{y_1, y_2, ..., y_i ..., y_k\}, \bigvee_{i=1}^{k} y_i \in Y\}$
- *χ-k-узел* к - элемент множества $\hat{Y}(k) = \{Y \times \sigma Y^k\}$
- ***χ-узел*** - элемент *множества* $\kappa \in \hat{Y} = \bigcup_{k=2}^{8} \hat{Y}(k)$
- *К-узлы графа* - связные подграфы $(\Gamma(v), \hat{e}v)$
  - *множество смежности* вершины *v* $\Gamma(v) = \hat{v}\hat{e}v$
  - Все узлы χ-графа $\mathbf{K}(X) = \{(\Gamma(\mathbf{v}), \hat{\mathbf{e}}\mathbf{v}) \mid \mathbf{v} \in V(X),\ d(v) > 1\}$

$\mathbf{G} = (\mathbf{N}, \mathbf{N}^2)$ - бесконечный полный граф

$\mathbf{C} = \mathbf{C}(\mathbf{G})$ - множество всех цепей

$\mathbf{K} = \mathbf{K}(\mathbf{G})$ - множество всех узлов

Отображения графов во множества разметок (сочетаний меток)

- Теорема. $(\exists \mu_c : \mathbf{C} \to \tilde{Y}) \wedge (\exists \mu_к : \mathbf{K} \to \hat{Y})$, т.е. множество цепей произвольного χ-графа X однозначно отображается в подмножество $\tilde{Y}(X)$ множества χ-цепей $\tilde{Y}$, а множество всех k-узлов X – в подмножество $\hat{Y}(X)$ множества χ-узлов $\hat{Y}$
  - Иначе говоря, существование функции разметки μ_V обуславливает существование функций $\mu_к : \mathbf{K} \to \hat{Y}$ и $\mu_c : \mathbf{C} \to \tilde{Y}$
  - Следствие $\exists \mu_c^{-1} \Rightarrow \bigcup_{v \in V(X)} \hat{cv} = \bigcup_{\alpha \in \tilde{Y}(X)} \mu_c^{-1}(\alpha) = X$
  - Следствие $\exists \mu_к^{-1} \Rightarrow \bigcup_{к \in \hat{Y}(X)} \mu_k^{-1}(к) = X$
- Существование обратных отображений μ_c^{-1} $\mu_к^{-1}$ - необходимо для установления изоморфизма хемографов посредством установления полноты их инвариантов

Инварианты на основе сочетаний меток χ-графов

- Будем говорить, что *χ-цепь* $\alpha \in \tilde{Y}$ *входит в хемограф* X, $\alpha \overline{\in} X$, если $\alpha \in \tilde{Y}(X)$.
- Соответственно, *вхождение χ-узла* $\kappa \in \hat{Y}$ *в* X ($\kappa \overline{\in} X$) соответствует $\kappa \in \hat{Y}(X)$.
- *Теорема. Необходимые условия изоморфизма двух хемографов*

$$\tilde{Y}(X_1) = \tilde{Y}(X_2) \qquad \hat{Y}(X_1) = \hat{Y}(X_2)$$

– Элементарные χ-инварианты графа X –

- *Булевы* $(\alpha \overline{\in} X)$ $(\kappa \overline{\in} X)$
- *Числовые* $|\{x \in \Pi(X) | \alpha \overline{\in} x\}|$ $|\{x \in \Pi(X) | \kappa \overline{\in} x\}|$

Критерии полноты кортеж-χ-инвариантов

- *χ-инвариантами* будем называть инварианты хемографов, основанные на отношениях вхождения χ-цепей и χ-узлов в хемографы.
- *Оператор формирования χ-кортеж-инварианта* $\hat{\imath}\boldsymbol{\iota}_{\mathbf{e}}$ по множеству элементарных инвариантов $\boldsymbol{\iota}_{\mathbf{e}}$ - $\hat{\imath}\boldsymbol{\iota}_{\mathbf{e}} = (\iota_j, \iota_k, \ldots, \iota_l),\ \ \iota_j, \iota_k, \ldots, \iota_l \in \boldsymbol{\iota}_{\mathbf{e}}$
  - Значение i-го элемента кортежа $\hat{\imath}[i]\boldsymbol{\iota}_{\mathbf{e}}(G) = \iota(G) \,|\, \underbrace{\lambda(\iota) = i}$
- *условие полноты инварианта*

  $\forall a \in \boldsymbol{\Gamma} : b \in \mathbf{I}(G) \Leftrightarrow \iota(a) = \iota(b)$

Функция нумерации элементарных инвариантов

- ***Теорема.*** *Кортеж-инвариант является полным инвариантом тогда и только тогда, когда для произвольной пары неизоморфных графов в соответствующих кортежах значений данного инварианта существует различающийся элемент.*

  – ***критерий полноты кортеж-инвариантов***

$$\forall a, b \in \boldsymbol{\Gamma} : \mathbf{I}(a) \cap \mathbf{I}(b) = \varnothing \Leftrightarrow \underset{i=1..n}{\exists}\ i : \hat{\imath}[i]\boldsymbol{\iota}_{\mathbf{e}}(a) \neq \hat{\imath}[i]\boldsymbol{\iota}_{\mathbf{e}}(b)$$

Комбинаторные оценки локальной полноты инвариантов

- ***Теорема 16.*** *Пусть задано множество прецедентов графов* Pr, *(*$\mathrm{Pr} \subset \mathbf{\Gamma}$*), множество χ-инвариантов* $\boldsymbol{\chi}$ *и кортеж-инвариант* $\iota = \hat{\imath}\boldsymbol{\chi}$ *. Тогда следующие утверждения эквивалентны:*
  - ι - *локально полный (над* Pr*),*
  - $\boldsymbol{\chi}$ *обеспечивает разрешимость соответствующей задачи над Pr при непротиворечивых метках изоморфности,*

1

$$\bigvee_{a,b\in Pr} \mathbf{iso}(a) \neq \mathbf{iso}(b) \Leftrightarrow \bigexists_{i=1..|\boldsymbol{\chi}|} i : \hat{\imath}[i]\boldsymbol{\chi}(a) \neq \hat{\imath}[i]\boldsymbol{\chi}(b)$$

$$\bigvee_{a,b\in Pr} \mathbf{iso}(a) \neq \mathbf{iso}(b) \Leftrightarrow \bigexists_{i=1..|\boldsymbol{\pi}'|} i : \hat{\imath}[i]\hat{\boldsymbol{\beta}}[a]\boldsymbol{\pi}' \neq \hat{\imath}[i]\hat{\boldsymbol{\beta}}[b]\boldsymbol{\pi}'$$

Для инвариантов на основе цепей

  - *над регулярным Pr,* $\boldsymbol{\chi}$ *обеспечивает регулярность*
  - *Задача распознавания изоморфных графов разрешима тогда и только тогда, когда для каждого графа из Pr разность класса изомерных по* ι *графов и класса изоморфных графов пуста.*

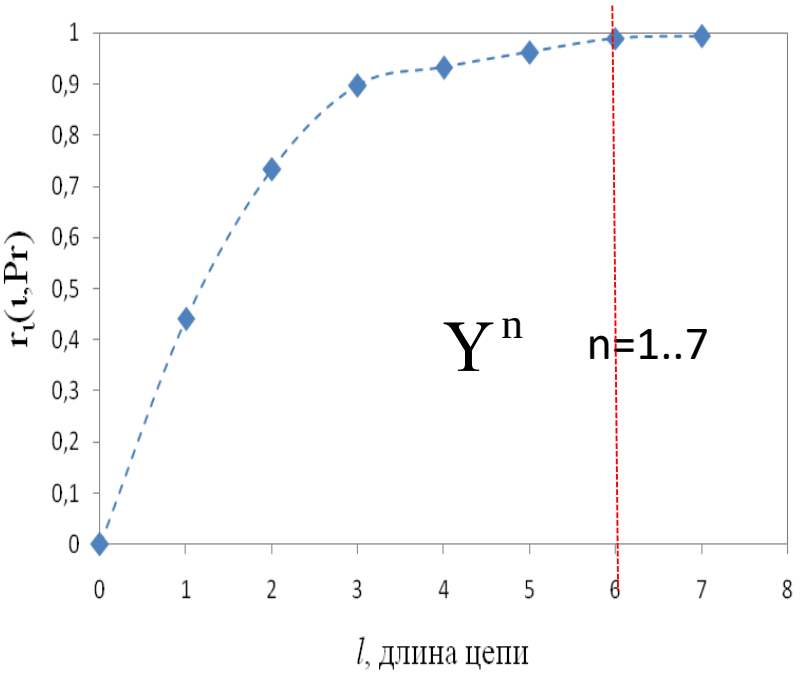
$$\bigvee_{a\in Pr} \mathbf{i\mu}(a,\iota,Pr) \setminus \mathbf{i}(a,Pr) = \varnothing$$

    - *локальное множество изоморфных графов* $\mathbf{i}(G,Pr) = \{g \in Pr \mid iso(g) = iso(G)\}$
    - *локальное множество изомерных графов* $\mathbf{i\mu}(G,\iota,Pr) = \{g \in Pr \mid \iota(g) = \iota(G)\}$
  - *Пусть* $r_\iota(\iota,Pr) = 1 - \frac{1}{|Pr|^2}\sum_{a\in Pr} |\mathbf{i\mu}(a,\iota,Pr) \setminus \mathbf{i}(a,Pr)|$*. Тогда* ***инвариант полон тогда и только тогда, когда*** $r_\iota(\iota,Pr) = 1$ *.*

$\mathbf{r}_\iota$ - комбинаторная оценка локальной полноты кортеж-инварианта

Тестирование локальной полноты для выбора оптимальных значений параметров разметок

- Тестирование
  - множество меток Y={H, C_{sp3}, C_{sp2}, C_{sp1}, N_{sp3}, N_{sp2}, N_{sp1}, O_{sp3}, O_{sp2}, P_{sp3}, S_{sp3}, X_{sp}, X_{sp2}, A}
  - 500000 структур PUBCHEM
  - семейства кортеж-инвариантов (n=1..7, k=3, 4)
    - Булевы инварианты над множеством χ-цепей
    - Булевы инварианты над множеством χ-узлов из k вершин



Расчет оценки локальной полноты $r_t(t, Pr)$ для кортеж-инвариантов над χ-цепями фиксированной длины

- При n=7 оценка полноты инвариантов достигла значения 0.99±0.01.
- при сравнительно коротких длинах χ-цепей (n=5) - 0.96 ±0.02.

Основы топологической теории анализа данных

2

- В рамках алгебраического подхода определяются
  - множества *начальных* (I_i) и *конечных информаций* (I_f)
  - множества прецедентов $Pr \subset I_i \times I_f$
  - исследуемые алгоритмы $A(\theta): I_i \to I_f$ (θ - вектор внутренних параметров алгоритма), настраиваемые по множествам прецедентов.
  - Исследуются свойств *разрешимости/регулярности* задач над Pr и *корректности/полноты* моделей алгоритмов $\{A(\theta)\}$ применяются факторизационный и метрический подходы, включающие анализ свойств компактности метрических конфигураций.
  - **Топологическая теория анализа данных** позволяет проводить систематические исследования возможных определений множеств I_i и I_f и выбирать определения, максимально адекватные исследуемой задаче.

Основные определения

$\mathbf{X} = \{x_1, x_2, ..., x_\alpha, ..., x_{N_0}\}$ – множество исходных описаний объектов $\mathbf{X} \subseteq S$

пространство допустимых **исходных** описаний объектов в проблемной области

J_{ob} - пространство допустимых объектов $I_i \subseteq I_1 \times I_2 \times ... \times I_k \times ... \times I_n$

$J_{ob} \subseteq I_1 \times I_2 \times ... \times I_k \times ... \times I_{n+l}$ $\quad I_f \subseteq I_{n+1} \times I_{n+2} \times ... \times I_{n+l}$

$I_k = \{\lambda_{k_1}, \lambda_{k_2}, ..., \lambda_{k_b}, ..., \lambda_{k_{|I_k|-1}}, \Delta\}$ - множество всех возможных значений k-ой компоненты формального признакового описания, k=1,…,*n+l*,

Δ - неопределённость, *n* – число признаков, *l* – число таргетных (прогнозируемых) переменных

$D: S \to J_{ob}$ сопоставляет объекту $s \in S$ допустимое описание $D(s)$

Формализация задачи - переход от множества $\mathbf{X} \subseteq S$ к $Q \subseteq J_{ob}$

$D(x_\alpha) = (\Gamma_1(x_\alpha) \times ... \times \Gamma_k(x_\alpha) \times ... \times \Gamma_{n+l}(x_\alpha))_\Delta \quad \Gamma_k : S \to I_k$, k=1,…,*n+l*

$m_\alpha = (\Gamma_1(x_\alpha) \times ... \times \Gamma_n(x_\alpha))_\Delta$ - *вектор значений признаков* (начальных информаций)

$t_\alpha = (\Gamma_{n+1}(x_\alpha) \times ... \times \Gamma_{n+l}(x_\alpha))_\Delta$ - *вектор значений таргетных переменных* (конечных информаций)

$\varphi: 2^S \to 2^{J_{ob}} \quad \varphi(\mathbf{X}) = \{D(x_\alpha) \mid x_\alpha \in \mathbf{X}\}$ формализует $\mathbf{X} = \{x_\alpha\}$ во множество прецедентов $Q = \{q_i \mid q_i = (m_i, t_i)\} \subseteq I_i \times I_f, q_i[k] \in I_k$

Примем, что для каждой Γ_k определена обратная ей **функция полного прообраза значения** $\lambda_{k_b} \in I_k \quad \Gamma_k^{-1}(\lambda_{k_b}) \subseteq \mathbf{X} \quad D^{-1}(q) = \bigcap_{k=1,n} \Gamma_k^{-1}(q[k])$

Th. Если $\forall x \in \mathbf{X} : x = D^{-1}(D(x))$ то $\forall_{Q^2}(q_1, q_2) : D^{-1}(q_1) \neq D^{-1}(q_2)$ **условие регулярности X**

- Анализ хемографов

$C = (\mathbf{R})$ **M**(C)=(d_{ij}(C)) $\longrightarrow \mathbf{X} \subseteq S$

Если I_k=[0,1] то множество элементарных инвариантов $\iota_e = \hat{\beta}\hat{\mu}_c^{-1}\alpha \cup \hat{\beta}\hat{\mu}_\kappa^{-1}\kappa$

Если $I_k \subset \mathbf{N}$ то множество элементарных инвариантов $\iota_e = \hat{\eta}\hat{\mu}_c^{-1}\alpha \cup \hat{\eta}\hat{\mu}_\kappa^{-1}\kappa$

Тогда $D(X) = \hat{\imath}\iota_e(X) \quad \Gamma_k(X) = \hat{\imath}[k]\iota_e(X)$ при произвольной функции нумерации инвариантов $\lambda : \iota_e \to \mathbf{N}$

Топологический подход к анализу I_i и I_f

«clopen»

При определенной функции $\varphi: 2^S \to 2^{J_{ob}}$ над множеством **X** формируется набор полных прообразов всех значений $\lambda_{k_b} \in I_k$ - множество $U(\mathbf{X}) = \{\Gamma_k^{-1}(\lambda_{k_b})\}$

Множество U(**X**) - предбаза ***топологии*** T(**X**)={∅, I, $a \cup b, a \cap b : a,b \in U(\mathbf{X})$}

Используя теоретико-множественное отношение включения как отношение порядка, элементы топологии могут быть частично упорядочены в *решётку* L(T(**X**))

$$L(T(\mathbf{X})) = \{a \vee b, a \wedge b : a,b \in T(\mathbf{X}), (a \geq b) \text{ или } (a \leq b)\}$$

Теорема. При выполнении условия регулярности **решётка** L(T(**X**)) – булева, в которой представлены три фундаментальные разновидности признаков:

- **булевы** признаки (все объекты с данным признаком находятся в вершине решётки) $\Gamma_k^{-1}(1)$
- **категорные** признаки (проецируются в антицепи решётки) $a \subseteq L(T(\mathbf{X}))$ $\forall_a x \neq y : \neg(x \subseteq y) \wedge \neg(y \subseteq x)$
- **числовые** признаки (проецируются в цепи решётки) $c \subset L(T(\mathbf{X}))$ $\forall_c x \neq y : (x \subseteq y) \vee (y \subseteq x) \neq \emptyset$

В случае k-го числового признакового описания, k=1…*n+l*, из линейного порядка $\lambda_{k_{b-1}} \leq \lambda_{k_b} \leq \lambda_{k_{b+1}}$ следует существование последовательности множеств $\Gamma_k^{-1}(\lambda_{k_1})$ $\Gamma_k^{-1}(\lambda_{k_1}) \cup \Gamma_k^{-1}(\lambda_{k_2})$, …, $\bigcup_{\beta=1}^{b} \Gamma_k^{-1}(\lambda_{k_\beta})$, …, I образующих цепь $A_k(\mathbf{X})$ так что каждому значению λ_{k_b} соответствует $u(\lambda_{k_b}) = \bigcup_{\beta=1}^{b} \Gamma_k^{-1}(\lambda_{k_\beta})$

$u(\lambda_{k_b})$ разбивает цепь $A_k(\mathbf{X})$ на две подцепи: нижнюю и верхнюю

Таким образом, k-ая числовая целевая переменная представлена цепью $A_k(\mathbf{X})$

$\mathrm{cdf}(A_k(\mathbf{X})) = \{(\lambda_{k_b}, \left|\{q_i \mid q_i[k] \leq \lambda_{k_b}\}\right| / N)\}$ *эмпирическая функция распределения (э.ф.р.)*

О порождении проблемно-ориентированных метрик на пространстве хемографов методами топологического анализа данных

3

Решётка L(T(**X**)) устанавливает частичный порядок элементов множества U(**X**)

Это позволяет сопоставить решётке L(T(**X**)) изоморфное метрическое пр-во $\rho_L : L^2 \to R^+$

метрическое пространство значений признаков

На основе топологии T(**X**) также возможно введение пространства с метрикой $\rho_q : Q^2 \to R^+$

метрическое пространство объектов

Для определения метрик вводятся понятия *изотонной оценки* на решётке и *окрестности* элемента в топологии

- *Оценка на решётке* L – *функция* $v : L \to R^+$ $\forall a,b : v[a] + v[b] = v[a \vee_L b] + v[a \wedge_L b]$ $\forall a,b : a \le b \Rightarrow v[a] \le v[b]$ *Оценка изотонна*
- *Окрестность* u точки x в топологии $T(\mathbf{X})$ - произвольное множество $u \subset T(\mathbf{X})$ $x \in u$

  Окрестность u *отделяет* точку x от точки y $(x \in u) \neq (y \in u)$ $x, y \in \mathbf{X}$

функция $\rho(x,y) = v[x \vee y] - v[x \wedge y]$ - метрика ρ_L $M_L(L(T(\mathbf{X})), \rho_L)$ метрическое пространство

изотонная оценка

Пример: $v[x] = h[x]$ $\rho(x,y) = |x \Delta y|$

метрическое пространство $M_q[L(T(\mathbf{X}))](Q, \rho_q)$ $q \subset Q$

- одна из вершин первого слоя L(T(**X**))
- набор n вершин решётки, $k = 1,\ldots,n$ соответствующих подмножествам $\Gamma_k^{-1}(q[k])$

Поскольку объект $x = D^{-1}(q) \in \mathbf{X}$ входит в каждое из $\Gamma_k^{-1}(q[k]) \in U(\mathbf{X})$ то для любых $q_1, q_2 \subset Q$ метрику $\rho_q(q_1, q_2)$ целесообразно определить как функцию $f_q : R^n \to R^+$ от вектора расстояний между множествами $\Gamma_k^{-1}(q[k])$ $\to f_q((\rho_L(\Gamma_k^{-1}(q_1[k]), \Gamma_k^{-1}(q_2[k])))$

Аддитивный вариант $\rho_q(q_1, q_2) = S(\sum_k \omega_k s(\rho_L(\Gamma_k^{-1}(q_1[k]), \Gamma_k^{-1}(q_2[k]))))$ → *метрики Хэмминга, Минковского*

Неаддитивный вариант $f_q(A) = \sup_{a \in A} a$ $A = (\rho_L(x,y))$

$\rho_L(x,y) = |x \Delta y| / |Q|$ → *метрика Колмогорова*

Проблемно-ориентированная теория для анализа хемографов

В задачах анализа хемографов функция D(X) определяется как $\hat{\mathbf{u}}_{\mathbf{e}}(\mathrm{X})$, $\Gamma_k(\mathrm{X}) = \hat{\imath}[k]\mathbf{u}_{\mathbf{e}}(\mathrm{X})$, а инвариантами в множестве $\mathbf{u}_{\mathbf{e}}$ являются элементарные χ-инварианты $\hat{\eta}[\mathrm{X}]\hat{\mu}_{\mathrm{c}}^{-1}\alpha$, $\hat{\eta}[\mathrm{X}]\hat{\mu}_{\kappa}^{-1}\kappa$, $\hat{\beta}[\mathrm{X}]\hat{\mu}_{\mathrm{c}}^{-1}\alpha$, $\hat{\beta}[\mathrm{X}]\hat{\mu}_{\kappa}^{-1}\kappa$ и др. Пусть заданы $\boldsymbol{\alpha} \subseteq \tilde{\mathrm{Y}}$, $\boldsymbol{\kappa} \subseteq \hat{\mathrm{Y}}$, $\boldsymbol{\pi} = \hat{\boldsymbol{\mu}}_{\boldsymbol{\kappa}}^{-1}\boldsymbol{\kappa} \cup \hat{\boldsymbol{\mu}}_{\mathbf{c}}^{-1}\boldsymbol{\alpha}$, $|\boldsymbol{\pi}| = \mathrm{n}$. Пусть множество элементарных χ-инвариантов $\mathbf{u}_{e}(X) = \hat{\boldsymbol{\beta}}[\mathbf{X}]\boldsymbol{\pi} = \{\hat{\beta}[X]\pi_i \mid \pi_i \in \boldsymbol{\pi},\ i = 1..n\}$, так что $\mathrm{D(X)} = \hat{\imath}\mathbf{u}_{\mathbf{b}}(\mathrm{X}) = \hat{\imath}\hat{\boldsymbol{\beta}}[\mathbf{X}]\boldsymbol{\pi}$, тогда ρ_q можно определить, например, как взвешенную *метрику Хэмминга*:

$$\rho_{\mathrm{q}}(X_1, X_2) = \frac{1}{n}\sum_{k=1}^{n} \omega_k \hat{\imath}[k]\hat{\boldsymbol{\beta}}[X_1]\boldsymbol{\pi} \oplus \hat{\imath}[k]\hat{\boldsymbol{\beta}}[X_2]\boldsymbol{\pi}. \tag{4}$$

При задании $\mathbf{u}_{e}(X) = \hat{\boldsymbol{\eta}}[\mathbf{X}]\boldsymbol{\pi} = \{\hat{\eta}[X]\pi_i \mid \pi_i \in \boldsymbol{\pi},\ i = 1..n\}$, $\hat{\imath}\mathbf{u}_{\boldsymbol{\eta}}(\mathrm{X}) = \hat{\imath}\hat{\boldsymbol{\eta}}[\mathbf{X}]\boldsymbol{\pi}$, ρ_q можно определить как взвешенную *метрику Минковского*,

$$\rho_q(X_1, X_2) = \sqrt[p]{\sum_{k=1}^{n} \omega_k \left|\hat{\imath}[k]\hat{\boldsymbol{\eta}}[X_1]\boldsymbol{\pi} - \hat{\imath}[k]\hat{\boldsymbol{\eta}}[X_2]\boldsymbol{\pi}\right|^p}.$$

В рамках топологического анализа порождение синтетических признаков хемографов, более информативных относительно k-ой переменной, чем исходные признаковые описания, осуществляется по двум направлениям: на основании метрики ρ_L или метрики ρ_q. Заметим, что в обоих случаях необходимо определение метрики ρ_L, которая используется сама по себе или для вычисления метрики ρ_q.

Порождение метрик типа ρ_L

Область значений целевой k-ой переменной I_k, k=n+1…n+*l*, проецируется в цепь $A_k(\mathbf{X}) = \left\langle \Gamma_k^{-1}(\lambda_{k_1}),...,u(\lambda_{k_b}),...,I \right\rangle$ решётки L(T(**X**)), $u(\lambda_{k_b}) = \bigcup_{\beta=1}^{b} \Gamma_k^{-1}(\lambda_{k_\beta})$

Каждому из элементов $A_k(\mathbf{X})$ сопоставлено множество объектов $\Gamma_k^{-1}(\lambda_{k_b})$ соответствующее значению λ_{k_β} k-ой переменной.

Поиск алгоритмов прогнозирования k-ой переменной заключается в нахождении во множестве цепей **A**(**X**) решётки L(T(**X**)) таких, что наиболее близки к цепи A_k(**X**)

Определим расстояние между цепями a,b $a = \langle a_1,...,a_i,...,I \rangle$ $b = \langle b_1,...,b_j,...,I \rangle$ суммой расстояний между соответствующими элементами этих цепей

$$\rho_A(a,b) = \min\left(\sum_{i=1,|a|} \rho_L(a_i, \arg\min_{b_j \in b} \rho_L(a_i,b_j)), \sum_{i=1,|b|} \rho_L(b_j, \arg\min_{a_i \in a} \rho_L(b_j,a_i))\right)$$

Естественным ограничением, накладываемым на поиск таких цепей, является то, что искомые цепи должны формироваться только на основании подмножеств множества **X**, соответствующих векторам $m_\alpha(x_\alpha) = (\Gamma_1(x_\alpha) \times ... \times \Gamma_n(x_\alpha))_\Delta$

предбаза U(**X**) сформирована только для признаковых описаний с k=1…n, что обозначим как $U(\mathbf{X})_{1,n}$, со множеством цепей $A(\mathbf{X})_{1,n}$

Оптимальный алгоритм прогнозирования k-ой числовой переменной

$$aa = \arg\min_{a \in \mathbf{A}(\mathbf{X})_{1,n}} \rho_A(A_k(\mathbf{X}), a)$$

При $D(X) = \hat{u}_e(X)$ элементы цепи *aa* соответствуют различным комбинациям χ-инвариантов хемографов

Синтетические признаки хемографов на основании метрики ρ_q

Прогнозирование k-ой числовой переменной может быть реализовано в рамках *хемометрического анализа* как проблема **согласования значений некоторой "экспертной" метрики ρ_e и "признаковой" метрики** с весами ρ_q

$$\arg\min_{\{\omega_k\}} \sum_{m=1}^{|X|} \sum_{j\neq m}^{|X|} \left| \rho_q(\{\omega_k\}, X_m, X_j) - \rho_e(X_m, X_j) \right|$$

Практически важным частным случаем экспертной метрики ρ_e является «**одномерная**» метрика на основе скаляра, в качестве которого выступает прогнозируемая числовая величина (например, модуль разности).

Теорема. При использовании одномерной экспертной метрики тестирование условия хемометрического анализа для метрики ρ_q в форме Хэмминга соответствует аддитивной схеме учета признаков.

$$\arg\min_{\{\omega_k\}} \sum_{m\neq m_0}^{|X|} \left| \rho_q(\{\omega_k\}, X_{m_0}, X_m) - T_m \right|$$

$$\arg\min_{\{\omega_k\}} \frac{1}{n} \sum_{m\neq m_0}^{|X|} \left| \sum_{k=1}^{n} \omega_k \hat{\imath}[k] \hat{\beta}[X_{m_0}] \boldsymbol{\pi} \quad \hat{\imath}[k] \hat{\beta}[X_m] \boldsymbol{\pi} - T_m \right|$$

В результате анализа топологии T(X) и решётки L(T(X)) могут быть получены различные алгоритмы прогнозирования k-ой числовой переменной

Интерпретации в рамках КМ

$$\arg\min_{\{\omega_k\}} \frac{1}{n} \sum_{\mathrm{m}\neq m_0}^{|\mathrm{X}|} \left| \sum_{k=1}^{n} \omega_k \hat{\imath}[k]\hat{\boldsymbol{\beta}}[X_{m_0}]\boldsymbol{\pi} \oplus \hat{\imath}[k]\hat{\boldsymbol{\beta}}[X_m]\boldsymbol{\pi} - \mathrm{T_m} \right|$$

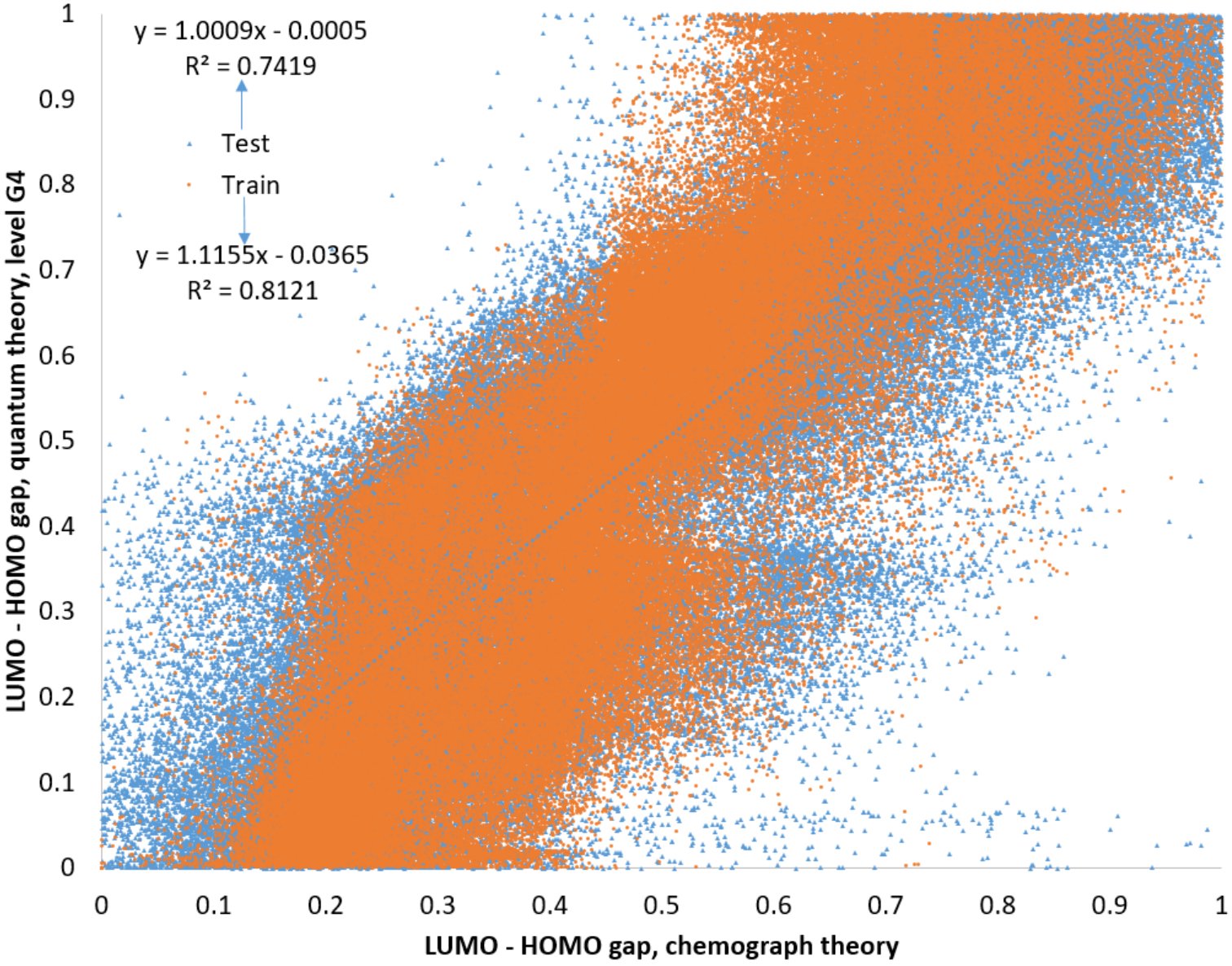
- **Теорема.** Результаты, получаемые, в рамках разработанного формализма, соответствуют
  - решению одноэлектронного уравнения Шредингера на фрагментах молекул с учётом перекрывания фрагментов
  - аддитивной схеме расчета электронной плотности в теории функционала плотности
  - учету интегралов перекрывания в теории молекулярных орбиталей (МО).

Апробация алгоритмов на выборке из 134000 молекул, кросс-валидационное тестирование

Конст.	**КМ-показатель**	**Единицы**	**r**	**r(c)**
A	Rotational constant A	GHz	0.77	0.73
B	Rotational constant B	GHz	0.74	0.73
C	Rotational constant C	GHz	0.72	0.71
M	Dipole moment	Debye	0.72	0.72
A	Isotropic polarizability	$Bohr^3$	0.69	0.67
HOMO	Energy of Highest occupied molecular orbital	Hartree	0.82	0.79
LUMO	Energy of Lowest occupied molecular orbital	Hartree	0.85	0.83
Gap	Gap difference between LUMO and HOMO	Hartree	0.86	0.83
r2	Electronic spatial extent	$Bohr^2$	0.67	0.67
ZPVE	Zero point vibrational energy	Hartree	0.85	0.85
U0	Internal energy at 0 K	Hartree	0.69	0.67
U	Internal energy at 298.15 K	Hartree	0.69	0.67
H	Enthalpy at 298.15 K	Hartree	0.69	0.67
G	Free energy at 298.15 K	Hartree	0.69	0.67
Cv	Heat capacity at 298.15 K	cal/M·K	0.75	0.75

Результаты кросс-валидационного тестирования алгоритмов имитационного моделирования для 15 КМ-показателей. Кросс-валидация включала 10 разбиений выборки «134К» на группы «случай-контроль» в соотношении 6:1. r, среднее значение рангового коэффициента корреляции на обучении; r(c), среднее значение рангового коэффициента корреляции на контроле.

Пример ранговой корреляции в координатах «алгоритм теории хемографов – алгоритм квантовой теории» для ширины щели LUMO-HOMO

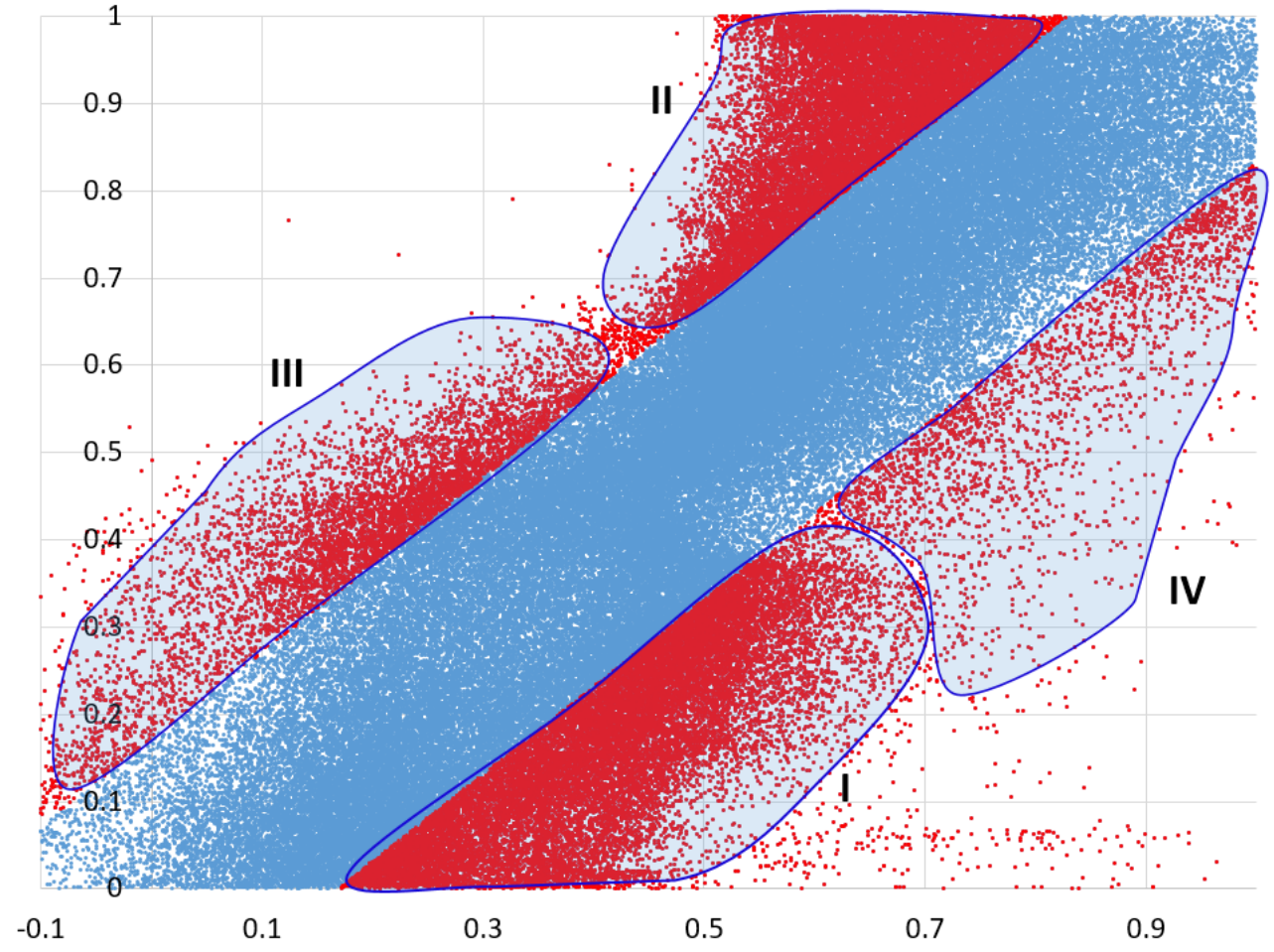


Интерпретируемость!

Chain	$\varphi_i(\hat{\imath}\chi, i, \mathrm{Pr})$	ω_i, arb. units
–^C(R,R1)^C(R2,R3)^CH_2^CH_2^CH_2–	0.0043	0.402
–HC(R)–H_2C–H_2C–H_2C–H_2C–	0.0063	0.369
–H_2C–H_2C–HC(R)–H_2C–CH_3	0.0077	0.337
–^CH_2–H^C*(R1)–H_2C–H^C*(R2)–O–	0.0042	0.333
–H^C*(R1)–H^C(R2)–O–H_2C–O–	0.0042	0.326
CH_3–HC(R)–H_2C–O–CH_3	0.0053	0.305
CH_3–H$C*(R1)–H$C*(R2)–O–$H_2$$C*–	0.0050	0.277
–^C(R1,R2)–$H_2$$C*–H$C*(R3)–$H_2$$C*–H^C(R)–	0.0046	0.277
NH_2–C(R1)=N–C(R2)=O	0.0047	–0.330
–$H_2$$C*–$C(R)=$C(R1)–H$C=H$C–	0.0095	–0.340
–C(R1)=HC–C(R2)=HC–HC=	0.0163	–0.340
–HC=C(R1)–N=C(R2)–HC=	0.0075	–0.341
–C(R1)=HC–HC=C(R2)–NH_2	0.0072	–0.425
NH_2–$C(R)=$C(R1)–HC=O	0.0065	–0.444
=$C(R)–N=$C(R1)–HC=O	0.0043	–0.463
–C(R1)=C(R2)–O–HC=C(R3)–	0.0045	–0.469
–C(R1)=C(R2)–HN–C(R3)=HC–	0.0051	–0.471

Примеры χ-цепей с наибольшими абсолютными значениями весов при расчёте щели LUMO-HOMO. Цепи расположены по убыванию весов . «^C» обозначает атом углерода со стерическим напряжением (например, в составе трёхатомного цикла); «C*» - хиральный центр типа «D» (как в D-аланине и т.п.); «$C», углерод в составе цикла; «R, R1, R2, R3» произвольные замещающие группы (т.н. «радикалы»).

Анализ ошибок расчетов LUMO-HOMO: регионы наиболее типичных ошибок.



Регион-I соответствует области достаточно низких значений щели LUMO-HOMO, которые были завышены (в среднем, на 0.15) при использовании «топологических» алгоритмов ИМ. Анализ атомного состава и структурных формул хемографов, попавших в регион-I, показал, что по сравнению с хемографами из региона-0, в этих хемографах в 2 раза встречались гидроксильные группы (функциональная группа «-OH»), в 10 раз чаще – **производные аммония , в 11 раз чаще - альдегиды (группа «-CHO»)**

Регион-II соответствует занижению (на -0.20) высоких значений щели LUMO-HOMO (0.7-1.0). В хемографах региона II двойные связи встречались в 5-10 раз реже, чем в остальных регионах, а атом фтора – в 2 раза чаще. Таким образом, в регионе-II гораздо чаще встречались **алифатические (насыщенные) фтор-содержащие соединения**.

Регион-III противоположен региону-I и соответствует области достаточно низких значений щели LUMO-HOMO, которые были занижены использованными алгоритмами ИМ. В регионе III тройная связь встречалась в 4-6 раз реже, чем в других регионах ошибок, а углероды с гибридизацией sp2 и не более, чем одним атомом водорода – в 4-10 раз чаще. При этом, цепи атомов углерода, соответствующие π-системам (C=C-C=C и др.), встречались в 3-8 раз чаще, а алифатические цепи (C-C-C и др.), наоборот, в 2-5 раз реже. Таким образом, среди точек в регионе III преобладали хемографы, описывающие **ароматические и другие π-системы**

Регион-IV противоположен региону-II и соответствует завышению средних и высоких значений щели LUMO-HOMO (0.5-0.8). Молекулы, содержащие 3 и более циклов, встречались в 3-10 раз чаще, чем в других регионах. Таким образом, ошибки в регионе-IV связаны с **преобладанием полициклических соединений**.

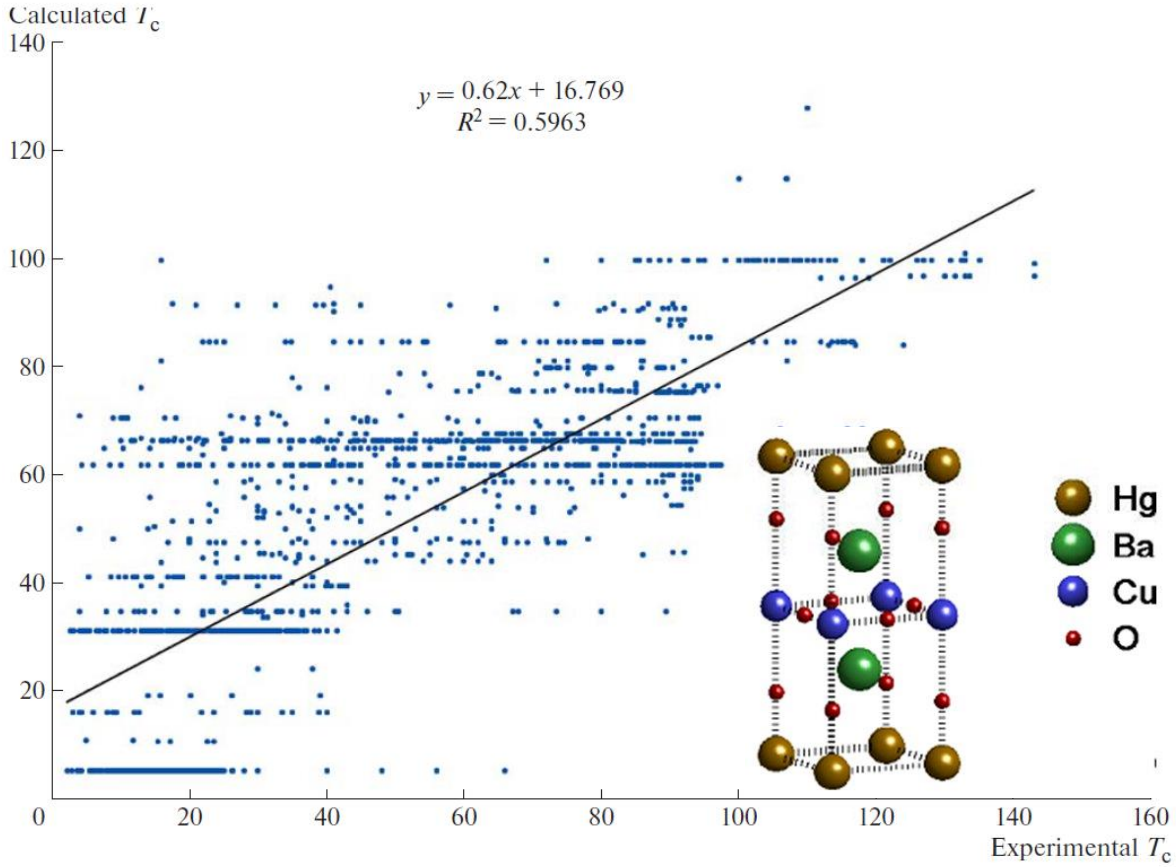
Экспертный анализ ошибок

- Информация о перечисленных выше 4 «ошибочных» классах молекул может учитываться при разработке более эффективных алгоритмов.
- **Пример**: разделение всей выборки 134К на две подгруппы – специфические хемографы (соответствуют полициклическим, ароматическим, алифатическим соединениям, n=26765) и все остальные хемографы
- **Результат**: коэффициент корреляции на контроле возрос от r(c)=0.83 до r(c)=0.88 при снижении стандартного отклонения от 0.17 до 0.11.

ПРАКТИКА: Полуколичественное прогнозирование Tc купратных сверхпроводников

Table 1. Examples of atomic chains with maximum weights in a linear model for calculating T_c. The chains are arranged in decreasing order of weights ω_i.

Chain	$\varphi_t(\hat{\imath}\chi, i, \mathrm{Pr})$	ω_i, K
Ba–O–Ca–O	0.108	7.43
Bi–Cu–O–O	0.048	6.49
Pb–O–O–Sr	0.048	6.31
Hg–O–Ba–O	0.036	6.08
Bi–O–O–Ca	0.008	5.92
Cu–O–Ca–O	0.607	5.49
Cu–O–Ba–O	0.382	5.27
Y–O–Cu–O	0.575	5.04
Pb–O–Cu–O	0.061	4.92
Y–O–Y–O	0.595	4.72
Eu–O–Eu–O	0.016	4.33
Y–O–Ni–O	0.011	4.3
Bi–Sr–O–O	0.178	4.16
Bi–O–Bi–O	0.269	4.11
Cu–Ba–O–O	0.047	−1.42
Ba–Ba–O–O	0.045	−1.48
Hg–O–Hg–O	0.019	−1.77
Pb–O–Sr–O	0.085	−1.83
Tl–O–Ba–O	0.013	−2.11



I. Yu. Torshin, K. V. Rudakov. Topological Data Analysis in Materials Science: The Case of High-Temperature Cuprate Superconductors. Patt Rec Image Analysis, 2020, Vol. 30, No. 2, pp. 262–274. DOI: 10.1134/S1054661820020157.

I. Yu. Torshin, V. A. Alyoshin, and E. V. Antipov, “Synthesis and properties of the high-temperature superconductor HgBa2CuO4+d,” Sverkhprovodimost: Fiz.,Khim., Tekh. 7 (10-12), 1579–1587 (1994).

S. N. Putilin, E. V. Antipov, O. Chmaissem, and M. Marezio, “Superconductivity at 94 K in HgBa2Cu04+d,” Nature, 362, 226–228 (1993).

ПРАКТИКА: Хемореактомный скрининг воздействия 2700 фармакологических препаратов на SARS-CoV-2 и виром человека как информационная основа для принятия решений по фармакотерапии COVID-19

Результаты. Установлены 62 препарата и 20 микронутриентов, которые характеризуются выраженным противовирусным действием в сочетании с минимальными побочными эффектами. Сопоставление полученных результатов с данными фундаментальных и клинических исследований показало, что для 31 из 62 препаратов имеются независимые подтверждения целесообразности их использования для лечения COVID-19. Установленные препараты являются ингибиторами коронавирусных белков и/или молекулами-адаптогенами, улучшающими функционирование клеток в условиях стресса при вирусной инфекции. Среди изученных «антикоронавирусных» микронутриентов наилучшим профилем безопасности, в т.ч. минимальным воздействием на виром здоровых людей, обладал глюкозамина сульфат.

Заключение. Перепрофилирование лекарственных препаратов, зарегистрированных в АТХ, может существенно ускорить нахождение более эффективных и безопасных подходов к фармакотерапии COVID-19. Перспективно применение ряда микронутриентов в программах долговременной профилактики коронавирусной инфекции, особенно у пожилых.

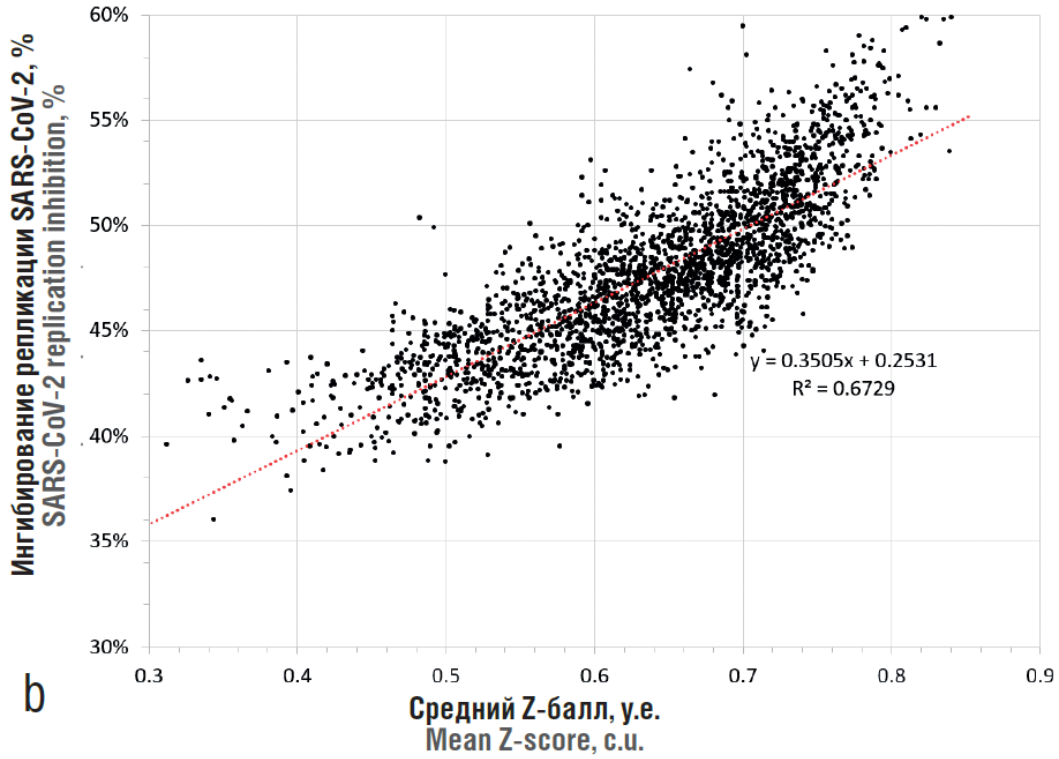


Таблица 1. Препараты, для которых продемонстрирована перспективность применения для лечения COVID-19
Table 1. Pharmaceuticals that have shown promise with respect to COVID-19 treatment

Соединение Compound	Z1	Z2	Ингибирование репликаций вируса, % Viral replication inhibition, %	Отказы, % Refusal, %	Класс по АТХ ATC class	Механизмы действия при COVID-19 Mechanism of action in case of COVID-19
Финголимод Fingolimod	0,84	0,66	55	1,73	L04AA Селективные иммунодепрессанты / L04AA Selective immunosuppressants	Модуляция рецепторов сфингозин-1-фосфата / Modulates sphingosine-1-phosphate receptors
Аргатробан Argatroban	0,79	0,7	52	5,17	B01AE Прямые ингибиторы тромбина / B01AE Direct thrombin inhibitors	Ингибирование тромбина / Inhibits thrombin
Паромомицин Paromomycin	0,71	0,75	49	2,01	A07AA Антибиотики / A07AA Antibiotics	Ингибитор основной протеазы SARS-CoV2 / Inhibits SARS-CoV2 main protease
Азитромицин Azithromycin	0,71	0,73	50	2,11	S01AA Антибиотики / S01AA Antibiotics	Ингибирование проникновения вирусов в клетку через CD147 TMPRSS2, противовоспалительное действие / Inhibits viral entry via CD147 TMPRSS2; anti-inflammatory effect
Маравирок Maraviroc	0,67	0,77	52	5,01	J05AX Другие противовирусные препараты / J05AX Other antivirals	Ингибитор основной протеазы и s-белка SARS-CoV-2 / SARS-CoV2 main protease and s-protein inhibitor
Налоксегол Naloxegol	0,81	0,63	46	2,72	A06AH Антагонисты периферических опиоидных рецепторов / A06AH Antagonists of peripheral opioid receptors	Блокирует связывание SARS-CoV-2 с АПФ / Blocks SARS-CoV-2 and ACE binding
Вальпроевая кислота Valproic acid	0,66	0,78	53	3,66	N03AG Производные жирных кислот / N03AG Fatty acid derivatives	Иммуномодуляция / Immunomodulation
Мемантин Memantine	0,78	0,65	47	0,56	N06DX Другие препараты от деменции / N06DX Other drugs for dementia	Ингибирование E-белка SARS-CoV-2 / Inhibits SARS-CoV-2 E-protein

Бромгексин Bromhexine	0,73	0,7	46	3,41	R05CB Муколитики / R05CB Mucolytics	Ингибитор протеазы TMPRSS2 / Inhibits TMPRSS2 protease
Амброксол Ambroxol	0,73	0,7	46	3,54	R05CB Муколитики / R05CB Mucolytics	Подавляет взаимодействие спайк-белка коронавируса с АПФ / Suppresses coronavirus spike protein and ACE interaction
Фавипиравир Favipiravir	0,79	0,63	46	7,29	J05AX Прочие противовирусные препараты / J05AX Other antivirals	Ингибирование репликации / Inhibits replication
Дапаглифлозин Dapagliflozin	0,75	0,67	52	2,92	A10BK Ингибиторы котранспортера натрия/глюкозы SGLT2 / A10BK Inhibitors of SGLT2 (sodium-glucose cotransporter 2)	Цитопротекция / Cytoprotection
Дисульфирам Disulfiram	0,73	0,68	49	3,78	P03AA Серосодержащие продукты / P03AA Sulfur-containing products	Ингибитор основной протеазы SARS-CoV-2 / Inhibits SARS-CoV-2 main protease
Метилпреднизолон Methylprednisolone	0,73	0,67	49	5,92	D07AC Кортикостероиды сильнодействующие (группа III) / D07AC Strong corticosteroids (Group III)	Противовоспалительное действие, повышение оксигенации / Anti-inflammatory; improves oxygenation

Выводы

- Предлагаемые процедуры оценочного имитационного моделирования свойств молекул находятся в русле, важном для решения задач теоретической и практической химии.
- В рамках КМ задаются наборы некоторых фундаментальных параметров отдельных частиц-волн (базисы орбиталей и др.) и параметры взаимодействия (например, корреляционые функционалы определенного вида). На основе заданных наборов параметров строятся всё более усложняющиеся схемы учета локальных взаимодействий между частицами с целью вычисления наблюдаемого свойства молекул («синтетический» подход). Затем, проводится сравнение с экспериментом.
- В рамках теории топологического анализа данных и парадигмы т.н. машинного обучения используется обратный подход («аналитический» подход): на основе известных из эксперимента наблюдаемых свойств молекул, с использованием вычислительных схем различной степени сложности (алгоритмов), строятся алфавиты атомов и словари локальных структур молекул, вычисляются вклады каждой из типов локальных структур молекулы в наблюдаемое свойство молекулы.

Образующие множества подграфов

Объединение множества подграфов П $\breve{\Pi} = \bigcup_{i=1}^{|\Pi|} \pi_i$

Множество всех замкнутых подграфов графа X(V,E)

$$\mathbf{\Pi}(X) = \left\{ (\mathbf{v},\mathbf{e}) \mid \mathbf{v} \subseteq V, \mathbf{e} \subseteq E, \overset{\mathbf{e}}{\forall} (v_1, v_2) : v_1 \in \mathbf{v}, v_2 \in \mathbf{v} \right\}$$

Множество подграфов $O \subseteq \mathbf{\Pi}(X)$ - ***образует*** X, если $\breve{O} = X$

(покрытие всех вершин и всех рёбер графа X)

Множество всех цепей хемографа X

$$\mathbf{C}(X) = \left\{ c = (V_c, E_c) \mid c \in \mathbf{\Pi}(X), \mid E_c \mid = \mid V_c \mid -1 > 0, \overset{|Vc|}{\underset{i=1}{\forall}} d(c, v_i) \le 2, \overset{|Vc|-1}{\underset{i=1}{\forall}} \overset{|Vc|}{\underset{j=i+1}{\forall}} v_i \ne v_j \right\}$$

Теорема 6. Множество всех цепей над χ-графом X образует X.

$$\breve{\mathbf{C}}(X) = X \rightarrow \exists \mathbf{c} \subseteq \mathbf{C}(X) : \breve{\mathbf{c}} = X$$

Теорема 10. Множество всех узлов χ-графа X образует X.

$$\breve{\mathbf{K}}(X) = X \rightarrow \exists \mathbf{к} \subseteq \mathbf{K}(X) : \breve{\mathbf{к}} = X$$

Тестирование локальной полноты для выбора оптимальных значений параметров разметок

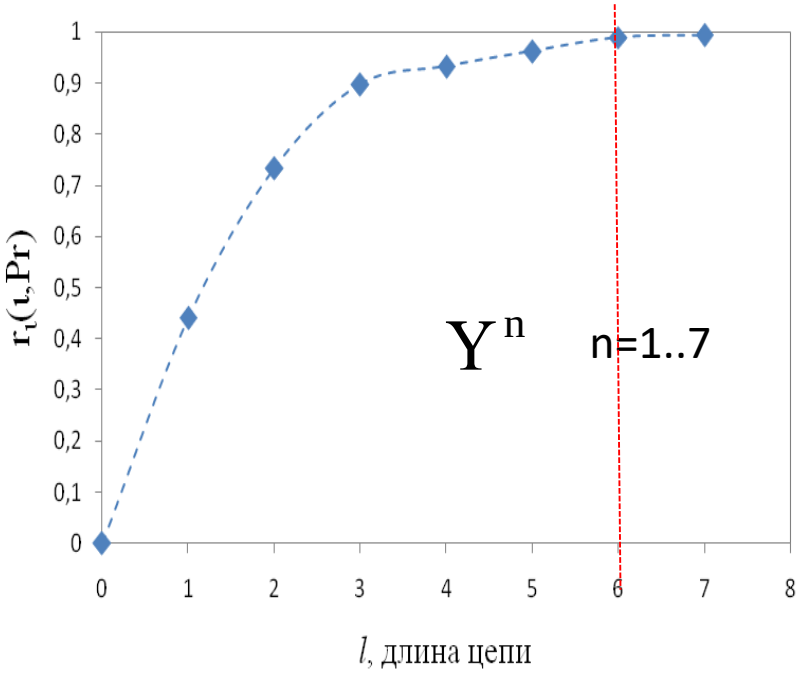
- Функции расстояния между хемографами
  - *метрика Хэмминга* над бинарными χ-инвариантами

$$d_{\chi b}(X_1, X_2) = \frac{1}{n}\sum_{i=1}^{n} \hat{\imath}[i]\hat{\boldsymbol{\beta}}[X_1]\boldsymbol{\pi} \oplus \hat{\imath}[i]\hat{\boldsymbol{\beta}}[X_2]\boldsymbol{\pi}$$

  - *метрики Минковского* над множеством численных χ-инвариантов

$$d_{\chi\eta}(X_1, X_2) = \sqrt[\frac{1}{p}]{\sum_{i=1}^{n} \left|\hat{\imath}[i]\hat{\boldsymbol{\eta}}[X_1]\boldsymbol{\pi} - \hat{\imath}[i]\hat{\boldsymbol{\eta}}[X_2]\boldsymbol{\pi}\right|^p}$$

- Тестирование
  - множество меток Y={H, C_{sp3}, C_{sp2}, C_{sp1}, N_{sp3}, N_{sp2}, N_{sp1}, O_{sp3}, O_{sp2}, P_{sp3}, S_{sp3}, X_{sp}, X_{sp2}, A}
  - 500000 структур PUBCHEM
  - семейства кортеж-инвариантов (n=1..7, k=3, 4)
    - Булевы инварианты над множеством χ-цепей
    - Булевы инварианты над множеством χ-узлов из k вершин



Расчет локальной полноты ($r_t(\iota, Pr)$) кортеж-инвариантов над χ-цепями фиксированной длины

- При n=7 оценка полноты инвариантов достигла значения 0.99±0.01.
- при сравнительно коротких длинах χ-цепей (n=5) - 0.96 ±0.02.

$$\rho_q(q_1, q_2) = S(\sum_k \omega_k s(\rho_L(\Gamma_k^{-1}(q_1[k]), \Gamma_k^{-1}(q_2[k]))))$$ **Взвешенные метрики**