ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2022. Т. 16. Вып. 2. С. 35–43

О ПРИМЕНЕНИИ ТОПОЛОГИЧЕСКОГО ПОДХОДА К АНАЛИЗУ ПЛОХО ФОРМАЛИЗУЕМЫХ ЗАДАЧ ДЛЯ ПОСТРОЕНИЯ АЛГОРИТМОВ ВИРТУАЛЬНОГО СКРИНИНГА КВАНТОВО-МЕХАНИЧЕСКИХ СВОЙСТВ ОРГАНИЧЕСКИХ МОЛЕКУЛ II: СОПОСТАВЛЕНИЕ ФОРМАЛИЗМА С КОНСТРУКТАМИ КВАНТОВОЙ МЕХАНИКИ И ЭКСПЕРИМЕНТАЛЬНАЯ АПРОБАЦИЯ ПРЕДЛОЖЕННЫХ АЛГОРИТМОВ*

И. Ю. Торшин[1]

Аннотация: Показаны соответствия между описаниями молекул в рамках теории хемографов, внутренними координатами молекул и ψ-функциями. Полученные результаты сопоставимы: (1) с решениями одноэлектронного уравнения Шредингера (УШ) на фрагментах молекул с учетом перекрывания фрагментов; (2) с аддитивными схемами расчета электронной плотности в теории функционала электронной плотности; (3) с учетом интегралов перекрывания в теории молекулярных орбиталей (МО). Апробация алгоритмов на выборке из 134 тыс. органических молекул показала ранговые корреляции порядка 0,75 (95%, достоверный интервал 0,67–0,85) между результатами расчетов по предлагаемым алгоритмам и значениями исследованных квантово-механических (КМ) показателей молекул. Скорость вычислений по предлагаемым алгоритмам на несколько порядков превышает скорость КМ вычислений, что важно для проведения скринингов молекул.

Ключевые слова: алгебраический подход; хемоинформатика; размеченные графы; комбинаторный анализ разрешимости

DOI: 10.14357/19922264220205

1 Введение

В первой части статьи [1] было показано, что при определении алфавита меток Y для произвольного хемографа $\mathrm{X} \in \mathbf{X}$ может быть вычислено множество $\tilde{\mathrm{Y}}(\mathrm{X}) \subset \tilde{\mathrm{Y}}$ всех χ-цепей в X, включающее все множества $\tilde{\mathrm{Y}}^m$ χ-цепей длины m и множество $\hat{\mathrm{Y}}(\mathrm{X}) \subset \hat{\mathrm{Y}}$ всех χ-узлов X. На основании множеств $\tilde{\mathrm{Y}}(\mathrm{X})$ и $\hat{\mathrm{Y}}(\mathrm{X})$ строятся соответствующие χ-инварианты и их кортежи посредством *оператора вхождения множества подграфов* $\boldsymbol{\pi} = \{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_n\} \subset \boldsymbol{\Gamma}$ в хемограф:

$$\hat{\beta}[X]\boldsymbol{\pi} = (|\boldsymbol{\pi} \cap \boldsymbol{\Pi}(X)| > 0)\,.$$

Обозначая через

$$\hat{\boldsymbol{\beta}}\boldsymbol{\pi}m = \left\{\hat{\beta}\boldsymbol{\pi}_1, \hat{\beta}\boldsymbol{\pi}_2, \ldots, \hat{\beta}\boldsymbol{\pi}_n\right\}$$

результат последовательного применения $\hat{\beta}$ к $\boldsymbol{\pi}$, а через $\hat{\imath}[i]\boldsymbol{\pi}(\mathrm{X})$ — i-й элемент кортеж-инварианта $\hat{\imath}\boldsymbol{\pi}$, получаем условие *хемометрического анализа*:

$$\arg\min_{\{\omega_k\}} \sum_{m=1}^{|\mathbf{X}|} \left| S\left(\sum_{k=1}^{n} \omega_k s\left(\hat{\imath}[k]\hat{\boldsymbol{\beta}}[X_m]\boldsymbol{\pi}\right)\right) - T_m \right|, \quad (1)$$

где T_m — значения прогнозируемой величины для молекул, соответствующих хемографам в обучающей выборке $\mathbf{X}$; $S, s : \mathrm{R} \to \mathrm{R}^+$ — «сглаживающие» функции. При $S \equiv 1$ и $s \equiv 1$ условие (1) соответствует определению взвешенной метрики Хэмминга

$$\rho_q(X_1, X_2) = \frac{1}{n}\sum_{k=1}^{n} \omega_k \hat{\imath}[k]\hat{\boldsymbol{\beta}}[X_1]\pi \otimes \hat{\imath}[k]\hat{\boldsymbol{\beta}}[X_2]\boldsymbol{\pi}. \quad (2)$$

В выражениях (1) и (2) настраиваемыми параметрами являются веса ω_k, а множество подграфов $\boldsymbol{\pi}$ и функции S и s в (1) задаются исследователем. Множество $\boldsymbol{\pi}$ определяется на основании *операторов построения прообраза* χ-цепи α, $\hat{\mu}_c^{-1}\alpha$, и *построения прообраза* χ-*узла* κ, $\hat{\mu}_\kappa^{-1}\kappa$, так что для $\boldsymbol{\alpha} \subset \tilde{\mathrm{Y}}$ определено $\hat{\boldsymbol{\mu}}_{\mathbf{c}}^{-1}\boldsymbol{\alpha} = \{\hat{\mu}_c^{-1}\alpha, \alpha \in \boldsymbol{\alpha}\}$, а для $\boldsymbol{\kappa} \subset \hat{\mathrm{Y}}$ — $\hat{\boldsymbol{\mu}}_\kappa^{-1}\boldsymbol{\kappa} = \{\hat{\mu}_\kappa^{-1}\kappa, \kappa \in \boldsymbol{\kappa}\}$. Тогда $\boldsymbol{\pi}$ опреде-

*Работа выполнена при поддержке РФФИ (проекты 19-07-00356, 18-07-00944 и 20-07-00537).

[1]Федеральный исследовательский центр «Информатика и управление» Российской академии наук, tiy135@yahoo.com

ляется как $\hat{\boldsymbol{\mu}}^1_{\mathbf{c}}\boldsymbol{\alpha}$, $\hat{\boldsymbol{\mu}}^1_{\boldsymbol{\kappa}}\boldsymbol{\kappa}$, $\hat{\boldsymbol{\mu}}^1_{\mathbf{c}}\boldsymbol{\alpha} \cup \hat{\boldsymbol{\mu}}^1_{\boldsymbol{\kappa}}\boldsymbol{\kappa}$ с использованием множеств χ-цепей фиксированной длины $m(\tilde{\mathrm{Y}}^m)$ и т. д.

Далее рассмотрены соответствия между результатами применения топологического анализа к хемографам и некоторыми математическими конструктам квантовой механики.

2 Основные постулаты квантовой механики

Аксиоматику, лежащую в основе математических конструкций КМ, удобно представлять в виде четырех постулатов [2], анализ которых позволяет рассмотреть соответствия между квантовой механикой и предлагаемым формализмом.

Постулат 1. *Состояние квантовой системы из N микрочастиц (электронов и ядер) полностью определяется волновой функцией от радиус-векторов частиц и времени (пси-функцией $\psi(\mathbf{x}, t) : \mathrm{R}^{3N+1} \to \mathrm{R}$), $\mathbf{x} \in \mathrm{R}^{3N}$ — вектор координат в конфигурационном пространстве, t — время. Квадрат ψ-функции отражает плотность вероятности состояния, заданного координатами частиц. В стационарном состоянии ψ-функция определяется как $\psi(\mathbf{x}) : \mathrm{R}^{3N} \to \mathrm{R}$.*

Постулат 2. *Наблюдаемая физическая величина A представима в виде линейного оператора $\hat{A}$, так что среднее значение A вычисляется как*

$$\bar{A} = \int\limits_{\mathrm{R}^{3N}} \psi^* \hat{A} \psi \, d\mathbf{x} \,,$$

где ψ^* комплексно сопряжена ψ.

Постулат 3. *Изменение волновой функции во времени определяется УШ.* Временна́я форма УШ для гамильтонана $\hat{\mathrm{H}}$ записывается как

$$\hat{\mathrm{H}}\psi = i\hbar \frac{\partial \psi}{\partial t} \,, \;\; \hat{\mathrm{H}} = \hat{\mathrm{T}} + \mathrm{V} \,.$$

Здесь $\hat{\mathrm{T}} = \sum \hat{T}_i$ — терм кинетической энергии, где

$$\hat{T}_i = \frac{\hbar^2}{2m_i} \Delta_i, \; \Delta_i = \frac{\partial}{\partial x_i^2} + \frac{\partial}{\partial y_i^2} + \frac{\partial}{\partial z_i^2}$$

(m_i — масса i-й частицы); $V(\mathbf{x}, t) : \mathrm{R}^{3N+1} \to \mathrm{R}$ — терм потенциальной энергии, где $\mathbf{x} \in \mathrm{R}^{3N}$ — конкатенация векторов координат ядер $\mathbf{R} = \left(\vec{\mathrm{R}}_\alpha\right)$ и электронов $\mathbf{r} = (\vec{\mathrm{r}}_i)$, $\vec{\mathrm{r}}_i = (x_i, y_i, z_i)$.

Постулат 4. *Электроны квантовой системы неразличимы.*

В гамильтоновых системах выполнено условие сохранения энергии, так что $V(\mathbf{x}) : \mathrm{R}^{3N} \to \mathrm{R}$ и $\psi(\mathbf{x}, t) = \psi(\mathbf{x})\boldsymbol{\chi}(t)$, что соответствует стационарной форме УШ:

$$\hat{\mathrm{H}}\psi(\mathbf{r}, \mathbf{R}) = E\psi(\mathbf{r}, \mathbf{R}) \,,$$

где E — полная энергия системы. После введения так называемого адиабатического приближения (Борна–Оппенгеймера: кинетическая энергия ядер пренебрежимо мала), гамильтониан УШ записывается как «электронный гамильтониан»:

$$\hat{\mathrm{H}}_e = \frac{1}{2} \sum_i \frac{\hbar^2}{2m_e} \Delta_i(\mathbf{r}) + \\ + V_{ee}(\mathbf{r}) + V_{en}(\mathbf{r}, \mathbf{R}) + V_{nn}(\mathbf{R}) \,, \quad (3)$$

где i-суммирование проводится по электронам (m_e); $V_{ee}(\mathbf{r}) = (1/2)\sum_i \sum_{j \neq i} 1/d_{ij}$, $d_{ij} = \|\vec{\mathbf{r}}_i - \vec{\mathbf{r}}_j\|$ — терм внутренней энергии межэлектронного взаимодействия; $V_{en} = -\sum_{i,\alpha} Z_\alpha / \mathbf{R}_{\alpha i}$, $\mathbf{R}_{\alpha i} = \|\vec{\mathbf{R}}_\alpha - \vec{\mathbf{r}}_i\|$ — терм электронно-ядерного, а $V_{nn} = \sum_{\alpha,\beta} Z_\alpha Z_\beta / \mathbf{R}_{\alpha\beta}$, $\mathbf{R}_{\alpha\beta} = \|\vec{\mathbf{R}}_\alpha - \vec{\mathbf{R}}_\beta\|$ — терм межъядерного взаимодействия (i, j-суммирование проводится по электронам, а α, β-суммирование — по ядрам). Вычисление КМ-показателей молекул, рассматриваемых в настоящей статье, основано на решениях уравнения (3).

После нумерации частиц в молекулярной системе $\mathrm{C} = (\mathbf{r}_j) \subset \mathrm{R}^3$, $j = 1, \ldots, N$, оператор $\hat{\mathrm{P}} : 2^{\mathrm{R}^3} \to \mathrm{R}^{3N}$ определяется как конкатенация $\hat{\mathrm{P}}\mathrm{C} = (\vec{\mathrm{r}}_1, \ldots, \vec{\mathrm{r}}_j, \ldots, \vec{\mathrm{r}}_N)$, так что $\psi(\hat{\mathrm{P}}\mathrm{C})$ зависит от декартовых координат. По построению, для $\hat{\mathrm{P}}$ всегда имеется обратный оператор $\hat{\mathrm{P}}^{-1}$, $\mathrm{C} \equiv \hat{\mathrm{P}}^{-1}\hat{\mathrm{P}}\mathrm{C}$. Декартовы координаты $\mathrm{C} = (\mathbf{r}_j)$ позволяют найти внутренние координаты $\mathbf{M}(\mathrm{C}) = (d_{ij}(\mathrm{C}))$ посредством такого $\hat{\mathrm{E}} : 2^{\mathrm{R}^3} \to \mathrm{R}^{\mathrm{C}^2_N}$, $\mathbf{M}(\mathrm{C}) = \hat{\mathrm{E}}\mathrm{C}$, что $d_{ij} = \|\vec{\mathrm{r}}_i - \vec{\mathrm{r}}_j\|$. Существует $\hat{\mathrm{E}}^{-1}$, так что $\mathbf{M}(\mathrm{C})$ позволяет найти C с точностью до афинного преобразования. Матрица инцидентности хемографа есть факторизация $\mathbf{M}(\mathrm{C})$ на основании правил теории химической связи [2, 3].

3 Интерпретация с точки зрения теории химической связи

Совокупность χ-инвариантов, в которые вовлечена данная вершина χ-графа, соответствующая одному из атомов молекулы, описывает некоторый локальный контекст данного атома в молекуле. В теории химической связи (композит классических и КМ-представлений) геометрия локального окружения атома описывается на основании гибридных состояний атомов, которые могут быть

использованы для порождения алфавита Y и словарей $\tilde{\mathrm{Y}}$ и $\hat{\mathrm{Y}}$ [3]. Присутствие в (1) весов ω_k соответствует суммированию свойства молекулы (k-я переменная) по χ-фрагментам (каждый из которых характеризуется определенным вкладом в это свойство), так что (1) подразумевает две гипотезы: (1) аддитивность и (2) фиксированность.

Определение 1. Гипотеза аддитивности: свойство всей молекулы представимо как сумма вкладов χ-инвариантов, каждый из которых соответствует определенному χ-фрагменту молекулы.

Определение 2. Гипотеза фиксированности вклада: каждый χ-инвариант вносит одинаковый вклад в исследуемое свойство во все молекулы, содержащие соответствующие χ-фрагменты.

4 Интерпретация в терминах одноэлектронной модели

Адиабатическое приближение (3) подразумевает, что терм $V_{nn}(\mathbf{R})$ фиксирован для заданной конфигурации ядер $\mathbf{R}$, так что $\mathbf{R}$ отражает параметры стационарного УШ. Одноэлектронное приближение упрощает (3) за счет пересмотра V_{ee}, которое аппроксимируется одноэлектронными операторами. В рамках одноэлектронного (Хартри–Фока) приближения $\psi(\mathbf{r},\mathbf{R})$ ищется в виде

$$\psi(\mathbf{r},\mathbf{R}) = \prod_i \psi_1\left(\vec{\mathrm{r}}_i,\mathbf{R}\right),$$

где каждая из $\psi_1(\vec{\mathrm{r}}_i,\mathbf{R})$ — решение задачи движения одного электрона в поле всех ядер, так называемого одноэлектронного УШ:

$$\left.\begin{array}{c} \hat{\mathrm{h}}_i\psi_{1,k}\left(\vec{\mathrm{r}}_i,\mathbf{R}\right) = e_{ik}\psi_{1,k}\left(\vec{\mathrm{r}}_i,\mathbf{R}\right)\,; \\ \hat{\mathrm{h}}_i = -\dfrac{\hbar^2}{2m_e}\,\Delta_i - \sum\limits_{\alpha}\dfrac{Z_\alpha}{\mathrm{R}_{\alpha t}} + V_i(i)\,; \\ \hat{\mathrm{H}}_e = \sum\limits_i \hat{\mathrm{h}}_i,\ \ \mathrm{E}_k = \sum\limits_i e_{ik}\,, \end{array}\right\} \qquad (4)$$

где $V_i(i) = (1/2)\sum_{i\neq j} 1/d_{ij}$ — терм потенциальной энергии, а $\psi_{1,k}$ ортонормированы [2]. В соответствии с постулатом 4 $\psi_{1,k}$ представляются в виде определителя, составленного из одноэлектронных линейно независимых функций.

Таким образом, в одноэлектронном приближении энергия системы представляется как сумма собственных значений фокианов $\hat{\mathrm{h}}_i$. Хотя в $\hat{\mathrm{h}}_i$ входят координаты всех ядер $\mathbf{R}$, с увеличением размера молекулы на движение электрона будут влиять только ближайшие ядра, что делает адекватным представление молекулы в виде набора локальных структурных фрагментов. Последнее соответствует суммированию свойств молекулы по χ-фрагментам в задаче (1), т. е. χ-фрагментам в предбазе $\mathbf{U}(\mathbf{X})$ сопоставлен определенный набор одноэлектронных волновых функций $\psi_{1,k}$, что отвечает гипотезе аддитивности. Выражение (1) усиливает ограничения на $\psi_{1,k}$: принимается, что любая $\psi_{1,k}$ одинакова в контексте структуры различных молекул.

5 Интерпретация в теории молекулярных орбиталей

В теории МО рассматриваются только орбитали валентных электронов молекулы. Молекулярные орбитали представляются линейными комбинациями (ЛК) атомных орбиталей (АО), что соответствует аппроксимации решения (4) водородоподобными ψ-функциями при больших (ангстремы) и малых (доли ангстрема) расстояниях [2]. Пусть произвольная МО Ψ представима как ЛК $\Psi = \sum_i c_i\psi_i$. Подставляя эту ЛК в выражение в постулате 2 и вводя условие нормировки Ψ для попарно неортогональных ψ_i, получаем

$$\overline{\mathrm{A}} = \frac{\sum_i\sum_j c_i c_j \mathrm{A}_{ij}}{\sum_i\sum_j c_i c_j \mathrm{S}_{ij}}\,,$$

где $\mathrm{A}_{ij} = \int_{\mathrm{R}^{3N}} \psi_i\hat{\mathrm{A}}\psi_j\,d\mathbf{x}$ и $\mathrm{S}_{ij} = \int_{\mathrm{R}^{3N}} \psi_i\psi_j\,d\mathbf{x}$ — так называемые интегралы перекрывания, характеризующие межэлектронные взаимодействия ψ_i. Если каждая из функций ψ_i в ЛК представляет ту или иную АО i-го атома молекулы, то матрица инцидентности $(m_{ij}(X))$ может рассматриваться как факторизация (S_{ij}), где бóльшим S_{ij} соответствуют бóльшие веса ребер m_{ij}.

Пусть имеется достаточно большое множество хемографов $\mathbf{X}$. Рассмотрим два фрагмента одной молекулы из $\mathbf{X}$, каждый из которых соответствует определенному χ-инварианту. Например, пусть $\alpha,\beta\in\tilde{Y}$, так что определены $\hat{\mu}_c^{-1}\alpha$ и $\hat{\mu}_c^{-1}\beta$ (случай с χ-узлами рассматривается аналогично). Определим три подмножества $\mathbf{X}$:

$$\begin{aligned} AB &= \left\{\mathrm{X}\in\mathbf{X} | \hat{\mu}_c^{-1}\alpha\overline{\in}\mathrm{X}, \hat{\mu}_c^{-1}\beta\overline{\in}\mathrm{X}\right\}; \\ A &= \left\{\mathrm{X}\in\mathbf{X} | \hat{\mu}_c^{-1}\alpha\overline{\in}\mathrm{X}\right\}\setminus AB; \\ B &= \left\{\mathrm{X}\in\mathbf{X} | \hat{\mu}_c^{-1}\beta\overline{\in}\mathrm{X}\right\}\setminus AB\,. \end{aligned}$$

Для определенности пусть $|\hat{\mu}_c^{-1}\alpha| = 1$ и $|\hat{\mu}_c^{-1}\beta| = 1$ (случаи с $|\hat{\mu}_c^{-1}\alpha| > 1$ и $|\hat{\mu}_c^{-1}\beta| > 1$ принципиально не отличаются).

В соответствии с определением 2 будем считать, что все χ-фрагменты $\hat{\mu}_c^{-1}\alpha$ описываются одной и той же ψ-функцией ψ_A, а все χ-фрагменты

$\hat{\mu}_c^{-1}\beta - \psi_B$. Взаимодействие между ψ_A и ψ_B зависит от расстояния d_{AB} между ними в каждой молекуле из AB: чем дальше расположены χ-фрагменты, тем меньше S_{AB} и тем более адекватно описание ψ_{AB} как ЛК ψ_A и ψ_B. Технически d_{AB} между χ-фрагментами в составе одной молекулы может оцениваться различными способами (среднее/минимальное расстояние между атомами, длина наикратчайшего пути и т. д.). С точки зрения теории МО подтверждением физического смысла гипотезы аддитивности является следующая теорема.

Теорема 1 (об аддитивной коррекции взаимодействий). *Пусть оценка взаимодействия между χ-фрагментами в произвольном хемографе монотонно убывает при увеличении расстояния между χ-фрагментами и не зависит ни от типов χ-фрагментов, ни от расположения χ-фрагментов в контексте хемографа, ни от выбранного способа измерения расстояния между фрагментами. Тогда вклады любых двух χ-фрагментов можно считать независимыми, а поправку на взаимодействие между парой χ-фрагментов учитывать как вклад третьего χ-фрагмента, образующего с парой χ-фрагментов связный подграф.*

Доказательство. Вне зависимости от процедуры порождения χ-фрагментов и способа измерения расстояния между χ-фрагментами, если для двух фрагментов молекулы не имеется третьего фрагмента, соединяющего их, то вклады обоих χ-фрагментов в общее свойство молекулы можно считать независимыми и перекрыванием таких удаленных орбиталей можно пренебречь ($\mathrm{S}_{AB} = 0$). Пусть два χ-фрагмента $\hat{\mu}_c^{-1}\alpha$ и $\hat{\mu}_c^{-1}\beta$, $|\hat{\mu}_c^{-1}\alpha| = 1$ и $|\hat{\mu}_c^{-1}\beta| = 1$, вносят вклады ω_α и ω_β в свойство произвольной молекулы, а $\omega_{\alpha\beta}$ — поправка на взаимодействие $\hat{\mu}_c^{-1}\alpha$ и $\hat{\mu}_c^{-1}\beta$ (случаю $|\hat{\mu}_c^{-1}\alpha| > 1$ соответствует $\omega_\alpha|\hat{\mu}_c^{-1}\alpha|$).

Пусть поправка на взаимодействие α и β — монотонно убывающая f^-, $\omega_{\alpha\beta} = f^-(d_{\alpha\beta})$, так что вклад α и β равен $\omega_\alpha + \omega_\beta + f^-(d_{\alpha\beta})$. Для m χ-фрагментов хемографа X вычислим расстояния $\{d_{ij}\}$ от i-го χ-фрагмента μ_i до всех остальных χ-фрагментов μ_j, $i \neq j$, так что для фрагментов с одинаковыми распределениями $\{d_{ij}\}$ суммарная поправка равна $s_i = \sum_{j\neq i} f^-(d_{ij})$.

Распределения расстояний $\{d_{ij}\}$ зависят от центральности расположения фрагмента. Центром графа будем считать i-е вершины с минимальными значениями «центральности» $c_i = \sum_{j=1,m} d_{ij}$. Для фрагментов с одинаковой центральностью c_i поправки s_i равны. Для двух фрагментов с разной центральностью пусть i_1 — более периферийный, а i_2 — более центральный, так что $c_{i_1} \geq c_{i_2}$ и $s_{i_1} \leq s_{i_2}$. Если имеется третий фрагмент i_3 из $\mathbf{\Pi}(\mathrm{X})$, соединяющий χ-фрагменты μ_{i_1} и μ_{i_2}, то $c_{i_1} \geq c_{i_3} \geq c_{i_2}$ и $s_{i_1} \leq s_{i_3} \leq s_{i_2}$, так что вклад всех трех фрагментов равен $\omega_{i_1} + \omega_{i_2} + \omega_{i_3} + f^-(d_{i_1 i_2}) + f^-(d_{i_1 i_3}) + f^-(d_{i_2 i_3})$. Однако по условию теоремы взаимодействие между χ-фрагментами не зависит от расположения χ-фрагментов в контексте хемографа, поэтому $s_{i_1} = s_{i_3} = s_{i_2} = s$. Тогда сумма поправок $\omega_{i_1 i_2}$ по всем парам χ-фрагментов μ_{i_1} и μ_{i_2} равна $\sum_{i=1,m} s_i$, т. е. является произведением числа χ-фрагментов хемографа m на константу s. Соответственно, вклад ω_{i_3} третьего фрагмента μ_{i_3} можно рассматривать как поправку на взаимодействие между двумя фрагментами μ_{i_1} и μ_{i_2}. Теорема доказана.

Следствие 1. Условию теоремы соответствуют наборы χ-фрагментов, полученные полным перебором χ-подграфов (χ-цепей или χ-узлов).

Следствие 2. Вычисление свойства всей молекулы осуществимо суммированием по χ-фрагментам.

Следствие 3. При учете взаимодействий χ-фрагментов по условию теоремы в вычисляемое свойство молекулы входит терм ms, равный произведению числа χ-фрагментов хемографа m на константу s.

Следствие 4. При выполнении условия теоремы свойство W молекулы X рассчитывается по аддитивной схеме

$$W = \sum_{i=1}^{m} \omega_i + ms\,,$$

т. е. суммированием по m χ-фрагментам, $\mu_i \in \in \mathbf{\Pi}(X)$. Из этого следует, что соответствующие ψ-функции ψ_i взаимно ортогональны (т. е. интегралы их перекрывания равны нулю).

Теорема 1 показывает, что при определенных условиях, накладываемых на процедуры построения множеств хемоинвариантов $\tilde{\mathrm{Y}}(\mathrm{X})$ и $\hat{\mathrm{Y}}(\mathrm{X})$, даже простейшая аддитивная схема расчета свойств молекулы, соответствующая постановке задачи хемореактивного анализа в простейшей форме (1), позволяет учитывать взаимодействия между χ-фрагментами, т. е. интегралы перекрывания S_{AB}. Более того, добавление в множества $\tilde{\mathrm{Y}}(\mathrm{X})/\hat{\mathrm{Y}}(\mathrm{X})$ χ-инвариантов, соответствующих «третьим» χ-фрагментам (т. е. тем, что образуют связный подграф для произвольной пары χ-фрагментов), позволяет предполагать ортогональность ψ-функций, соответствующих χ-инвариантам в таких $\tilde{\mathrm{Y}}(\mathrm{X})/\hat{\mathrm{Y}}(\mathrm{X})$ (следствие 4).

Экспериментальная верификация всего комплекса гипотез в условии теоремы 1 заключается: (1) в оценке корреляции между значениями КМ параметров молекул и результатами расчетов по предлагаемой в следствии 4 аддитивной схеме; (2) в оценке расстояний между соответствующими χ-фрагментами в множестве AB на основе эмпирической функции распределения (э. ф. р.) рассто-

яний $\{d_{AB}\}$ между χ-фрагментами $\hat{\mu}_c^{-1}\alpha$ и $\hat{\mu}_c^{-1}\beta$ с последующим анализом ρ-спектров и других свойств возникающих при этом ρ_L-конфигураций [1].

В целом, с точки зрения теории МО подразумевается делокализация электронов вокруг χ-фрагментов, соответствующих χ-цепям/χ-узлам. Обобществление электрона в случае χ-узла вполне представимо, так как последний представляет ближайшее ковалентное окружение произвольного атома.

6 Интерпретация с точки зрения теории функционала электронной плотности

Центральная идея теории функционала электронной плотности (density functional theory, DFT) заключается в переформулировке постулатов 1–4 в формах, включающих электронную плотность системы

$$\rho\left(\vec{\mathrm{r}}\right) = N \int\limits_{\mathrm{R}^{3N}} |\Psi(\mathbf{x})|^2 d\mathbf{x}\,.$$

Последняя имеет четкий физический смысл и может быть экспериментально оценена. Вводится такое преобразование F, что

$$\Psi(\mathbf{x}) = \mathrm{F}\left(\rho(\vec{\mathrm{r}})\right); \quad \bar{\mathrm{A}} = \left\langle \mathrm{F}\left(\rho(\vec{\mathrm{r}})\right) |\hat{\mathrm{A}}| \mathrm{F}\left(\rho(\vec{\mathrm{r}})\right) \right\rangle,$$

а полная энергия системы в приближении Борна–Оппенгеймера равна

$$\mathrm{E}\left(\rho(\vec{\mathrm{r}})\right) = \mathrm{F}_{\mathrm{HK}}\left(\rho(\vec{\mathrm{r}})\right) + V_{en}(\mathrm{F}\left(\rho(\vec{\mathrm{r}})\right);$$
$$\mathrm{F}_{\mathrm{HK}}\left(\rho(\vec{\mathrm{r}})\right) = \hat{\mathrm{T}}\mathrm{F}\left(\rho(\vec{\mathrm{r}})\right) + V_{ee}\left(\mathrm{F}\left(\rho(\vec{\mathrm{r}})\right)\right).$$

Процедуры вычисления $\rho(\vec{\mathrm{r}})$ естественным образом допускают аддитивные модели. Соответственно, представление молекулы как набора χ-фрагментов — это способ аддитивного, пофрагментного описания карты электронной плотности молекулы.

7 О применении алгоритмов для скрининга молекул

Очевидно, что точность вычислений по аддитивной схеме (1), вследствие весьма сильных предположений об аддитивности и постоянстве вклада χ-инвариантов (определения 1 и 2), вряд ли когда-нибудь приблизится к результатам, получаемым в рамках разработанных ранее вычислительных схем КМ. Тем не менее ряд особенностей корреляционного облака точек

$$\mathrm{O}(\mathbf{X}) = \{(\mathrm{W}_m(\mathrm{X}), \mathrm{T}_m(\mathrm{X}))\,, \quad \mathrm{X} \in \mathbf{X}\,,\ \ m = 1, \ldots, |\mathbf{X}|\}\,,$$

где $\mathrm{W}_m(\mathrm{X})$ вычисляется для всех $\mathrm{X} \in \mathbf{X}$ (например, в соответствии с решением задачи (1) или с использованием более сложных схем алгебраического подхода), позволяет оценить практическую применимость получаемых результатов.

Во-первых, кросс-валидационные оценки коэффициента корреляции и других функционалов оценки адекватности моделей на $\mathrm{O}(\mathbf{X})$ (стандартное отклонение, коэффициент детерминации, комбинаторные и прочие функционалы робастного линейного сглаживания, различные статистические функционалы и др.) позволяют разносторонне оценить качество оценочных расчетов величин T посредством модели W. Во-вторых, может быть оценена релевантность характеристик облака точек $\mathrm{O}(\mathbf{X})$ для решения соответствующих задач классификации и высокопроизводительного скрининга молекул *in silico*.

Следующая теорема может быть полезна для планирования вычислительных экспериментов и анализа полученных данных.

Теорема 2 (о скрининге). *Точность классификации хемографов из* $\mathbf{X}$ *по интересующим процентилям значений пропорциональна степени покрытия корреляционного облака точек* $\mathrm{O}(\mathbf{X}) = \{(\mathrm{W}_m, \mathrm{T}_m)\}$ *ячейками главной диагонали координатной сетки, образованной соответствующими процентилями значений* T и W.

Доказательство проводится посредством рассмотрения в решетке $\mathrm{L}(\mathrm{T}(\mathbf{X}))$ цепей $\mathrm{A}(\mathbf{X})$ и $\mathrm{A}'(\mathbf{X})$, соответствующих величине T и модели W. Рассматриваются э. ф. р. $\mathrm{cdf}(\mathrm{A}(\mathbf{X}))$ и $\mathrm{cdf}(\mathrm{A}'(\mathbf{X}))$, взаимно однозначное соответствие между процентилями $\Pi(p, \mathrm{cdf}(\mathrm{A}(\mathbf{X})))$ и $\Pi(p, \mathrm{cdf}(\mathrm{A}'(\mathbf{X})))$, так что на основе $\mathrm{O}(\mathbf{X})$ вычисляются проценты ошибок классификации 1-го и 2-го типа.

Следствие 1. Величина коэффициента корреляции $r(\mathrm{O}(\mathbf{X}))$ — непрямая характеристика аккуратности классификации по процентилям значений.

Следствие 2. Разность коэффициентов корреляции на обучении и контроле косвенно характеризует переобученность алгоритма классификации.

8 Результаты экспериментальной апробации

Тестирование моделей порождения информативных числовых признаков хемографов, основан-

Результаты кросс-валидационного тестирования разработанных скрининговых алгоритмов для 15 КМ показателей молекул; $r(c)$ — среднее значение рангового коэффициента корреляции на контроле, $SD(c)$ — стандартное отклонение в прогнозировании ранга КМ-показателя на контроле

Константа	КМ-показатель	Единицы	r	$r(c)$	$SD(c)$
A	Вращательная константа A	ГГц	0,77	0,73	0,18
B	Вращательная константа B	ГГц	0,74	0,73	0,19
C	Вращательная константа C	ГГц	0,72	0,71	0,20
M	Дипольный момент	Дебай	0,72	0,72	0,20
α	Изотропная поляризуемость	Бор^3	0,69	0,67	0,21
HOMO	Энергия высшей занятой МО	Хартри	0,82	0,79	0,17
LUMO	Энергия низшей вакантной МО	Хартри	0,85	0,83	0,15
Gap	Зазор LUMO-HOMO	Хартри	0,86	0,83	0,15
r_2	Электронный пространственный экстент	Бор^2	0,67	0,67	0,21
ZPVE	Вибрационная E нулевого уровня	Хартри	0,85	0,85	0,15
U_0	Внутренняя энергия (0 K)	Хартри	0,69	0,67	0,21
U	Внутренняя энергия (298,15 K)	Хартри	0,69	0,67	0,21
H	Энтальпия (298,15 K)	Хартри	0,69	0,67	0,21
G	Свободная энергия (298,15 K)	Хартри	0,69	0,67	0,21
C_v	Теплоемкость (298,15 K)	кал/(М·К)	0,75	0,75	0,19

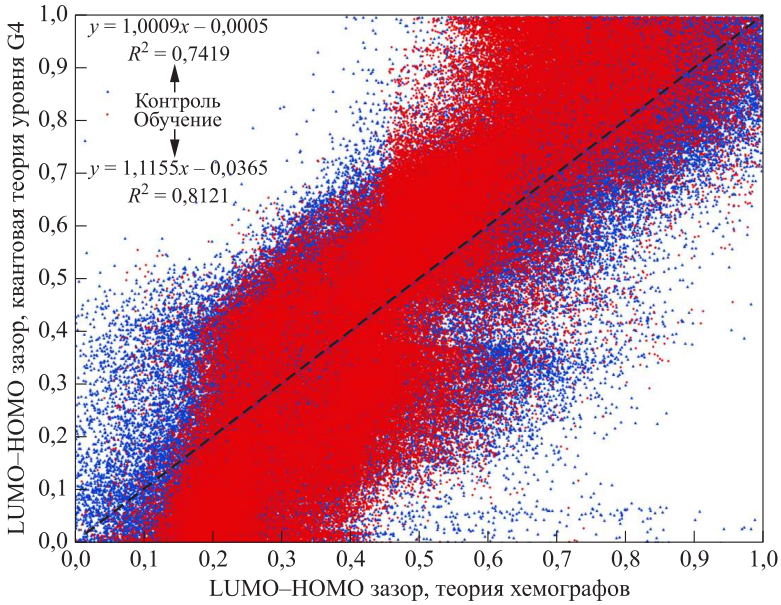
ных на решении задач типа (1) и соответствующих алгоритмов прогнозирования числовых переменных [4], было проведено на выборке из 134 000 стабильных органических молекул с максимум девятью тяжелыми атомами C, O, N и F (далее — 134K) [5]. Исходные описания хемографов в 134K представлены в виде матриц $\mathrm{M}(X)$, отражающих кратности химических связей. Было использовано множество меток Y, включающее элементы декартова произведения химического типа элемента на заряд и допустимые гибридизационные состояния атомов [6]. Над Y строились инварианты из семейств $\hat{\imath}\hat{\boldsymbol{\beta}}[\mathbf{X}]\hat{\boldsymbol{\mu}}_{\mathbf{c}}^{-1}\tilde{\mathrm{Y}}^n$ ($n = 1, \ldots, 7$), $\hat{\imath}\hat{\boldsymbol{\beta}}[\mathbf{X}]\hat{\boldsymbol{\mu}}_{\boldsymbol{\kappa}}^{-1}\hat{Y}(k)$ ($k = 3, \ldots, 7$) и $\hat{\imath}\hat{\boldsymbol{\beta}}[\mathbf{X}](\hat{\boldsymbol{\mu}}_{\mathbf{c}}^{-1}\tilde{\mathrm{Y}}^n \cup \hat{\boldsymbol{\mu}}_{\boldsymbol{\kappa}}^{-1}\hat{Y}(k))$. Результаты тестирования регулярности по Журавлеву [6] позволили найти оптимальные значения k и n ($k = 4$; $n = 5$). Как и в работе [4], прогнозирование числовых величин проводилось алгоритмически с *линейным распознающим оператором* на основании кортежинварианта $\hat{\imath}\boldsymbol{\iota}_{\mathbf{e}} = \hat{\imath}\hat{\boldsymbol{\beta}}[\mathbf{X}](\hat{\boldsymbol{\mu}}_{\mathbf{c}}^{-1}\tilde{\mathrm{Y}}^5 \cup \hat{\boldsymbol{\mu}}_{\boldsymbol{\kappa}}^{-1}\hat{Y}(4))$ и корректором из 6 операций, настраиваемыми мультистартовой стохастической оптимизацией.

Результаты серии кросс-валидационных экспериментов (10 разбиений 134K в соотношении «случай–контроль» 6 : 1) показали наилучшие результаты: (1) при учете эффектов атомов водорода; (2) при использовании линейного решающего правила и единичных «сглаживающих» функций $\mathrm{S}, \mathrm{s} : \mathrm{R} \to \mathrm{R}^+$; (3) при использовании коррекции на число χ-фрагментов хемографа (см. теорему 1). Результаты вычислительных экспериментов суммированы в таблице.

Несмотря на отличия в аккуратности скрининговых оценок различных КМ-свойств молекул (см. таблицу), алгоритмы для вычисления всех 15 свойств показали приемлемую обобщающую способность. Последняя может быть косвенно охарактеризована, например, различиями между значениями коэффициентов корреляции r и $r(c)$, полученных соответственно на обучении и контроле (теорема 2), которые составили в среднем всего 0,016 (95%, достоверный интервал 0,003–0,041). Одним из лучших скрининговых алгоритмов, разработанных в рамках топологической теории хемографов, оказался алгоритм вычислений ширины щели LUMO-HOMO (Lowest Unoccupied and Highest Occupied Molecular Orbitals): $r = 0{,}86$ на обучении и $r(c) = 0{,}83$ на контроле при стандартном отклонении 0,14–0,17 (см. рисунок). Несмотря на заметную «размытость» корреляционного облака $O(\mathbf{X})$ на рисунке, при скрининге молекул по LUMO-HOMO квартиль наибольших значений позволяет выделить 77% соединений с наибольшими значениями данного свойства молекул (теорема 2).

При экспертном анализе ошибок прогнозирования было установлено, что разделение всей выборки 134K на две подгруппы молекул: полициклические, ароматические и алифатические соединения ($n = 26\,765$) и все остальные — с раздельным обучением на каждой из подгрупп позволило улучшить кросс-валидационную корреляцию для зазора LUMO-HOMO от $r(c) = 0{,}83$ до 0,88 при снижении стандартного отклонения от 0,17 до 0,11.

Анализ весов ω_i и значений $\varphi_\iota(\hat{\imath}\boldsymbol{\chi}, i, \mathrm{Pr})$ (см. теорему 2 в [1]) χ-инвариантов позволяет выявить хемоинварианты, которые вносят наибольшие абсолютные вклады. Например, в увеличение щели LUMO-HOMO наибольший вклад вносили хемо-



Пример ранговой корреляции для ширины щели LUMO-HOMO

инварианты, содержащие *атомы углерода с выраженным стерическим напряжением*, *алифатические цепи*, а в сужение — π-*системы*, что полностью соответствует основам теории химической связи.

При условии проведения предварительной подготовки (конвертирование матрицы инцидентности каждого хемографа X в множество хемоинвариантов $\hat{\boldsymbol{\beta}}[\mathrm{X}](\hat{\boldsymbol{\mu}}_{\mathbf{c}}^{-1}\tilde{\mathrm{Y}}^{5}(\mathrm{X})) \cup \hat{\boldsymbol{\mu}}_{\boldsymbol{\kappa}}^{-1}\hat{Y}(4,\mathrm{X})) \subset \boldsymbol{\iota}_{\mathbf{e}}$, настройка вектора параметров $\theta(\mathrm{Pr})$) скорость вычислений алгоритма $\hat{\mathrm{A}}(\theta(\mathrm{Pr}))$ становится на несколько порядков выше, чем высокоточные КМ-расчеты. Таким образом, разработанные алгоритмы приемлемы для проведения виртуальных скринингов КМ-свойств молекул.

9 Заключение

Предлагаемые процедуры скринингового моделирования КМ-свойств молекул находятся в русле, важном для решения задач теоретической и практической химии. В теоретической химии крайне важна разработка моделей, приемлемых для всех классов соединений и позволяющих устанавливать полуколичественные взаимосвязи между электронно-пространственным строением молекул и их свойствами. Такие модели должны обеспечивать выделение структурных признаков, определяющих свойства молекул, возможные реакции молекул, прогнозировать эффекты модификации структуры молекулы. Необходимость таких моделей очевидна хотя бы потому, что число возможных органических молекул измеряется сотнями миллиардов и получить данные точных количественных КМ-расчетов для каждой из таких молекул не представляется возможным.

Предлагаемые в настоящей работе «топологические» модели отличает хорошая физико-химическая интерпретируемость (в том числе в терминах квантовой теории) и высокая скорость вычислений. Эти особенности разработанных моделей обеспечивают возможность их применения для решения широкого круга задач, таких как оценка КМ свойств метаболитов и олигопептидов, поиск/дизайн молекул с заданными наборами КМ-свойств в рамках решения задач материаловедения, дизайн новых лекарств, репозиционирование уже известных лекарств и др.

Литература

1. О применении топологического подхода к анализу плохо формализуемых задач для построения алгоритмов виртуального скрининга квантово-механических свойств органических молекул I: Основы проблемно ориентированной теории // Информатика и её применения, 2022. Т. 16. Вып. 1. С. 39–45.
2. *Степанов Н. Ф.* Квантовая механика и квантовая химия. — М.: Мир, 2001. 519 с.
3. *Torshin I. Yu., Rudakov K. V.* On the application of the combinatorial theory of solvability to the analysis of chemographs. Part 1: Fundamentals of modern chemical bonding theory and the concept of the chemograph // Pattern Recognition Image Analysis, 2014. Vol. 24. No. 1. P. 11–23.

4. *Torshin I. Y., Rudakov K. V.* On the procedures of generation of numerical features over partitions of sets of objects in the problem of predicting numerical target variables // Pattern Recognition Image Analysis, 2019. Vol. 29. No. 4. P. 654–667. doi: 10.1134/S1054661819040175.

5. *Ramakrishnan R., Dral P., Rupp M.* Quantum chemistry structures and properties of 134 kilo molecules // Scientific Data, 2014. Vol. 1. No. 1. Art. 140022. 7 p. doi: 10.1038/sdata.2014.22.

6. *Torshin I. Yu., Rudakov K. V.* On the application of the combinatorial theory of solvability to the analysis of chemographs. Part 2. Local completeness of invariants of chemographs in view of the combinatorial theory of solvability // Pattern Recognition Image Analysis, 2014. Vol. 24. No. 2. P. 196–208.

Поступила в редакцию 05.04.21

ON THE APPLICATION OF A TOPOLOGICAL APPROACH TO ANALYSIS OF POORLY FORMALIZED PROBLEMS FOR CONSTRUCTING ALGORITHMS FOR VIRTUAL SCREENING OF QUANTUM-MECHANICAL PROPERTIES OF ORGANIC MOLECULES II: COMPARISON OF FORMALISM WITH CONSTRUCTIONS OF QUANTUM MECHANICS AND EXPERIMENTAL APPROBATION OF THE PROPOSED ALGORITHMS

I. Yu. Torshin

Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation

Abstract: Correspondences between descriptions of molecules in the framework of the theory of chemographs, internal coordinates of molecules, and ψ-functions are shown. The results obtained are comparable: (i) with the solutions of the one-electron Schrödinger equation on fragments of molecules with allowance for the overlap of fragments; (ii) with additive schemes for calculating electron density in the electron density functional theory; and (iii) with allowance for overlap integrals in the theory of molecular orbitals. Approbation of the algorithms on a sample of 134 thousand organic molecules showed rank correlations of the order of 0.75 (95%, reliable interval 0.67–0.85) between the results of calculations using the proposed algorithms and the values of the investigated quantum mechanical properties of molecules. The calculation speed of the proposed algorithms is several orders of magnitude higher than the speed of quantum mechanical calculations which is important for screening the molecules.

Keywords: algebraic approach; chemoinformatics; labeled graphs; combinatorial solvability analysis

DOI: 10.14357/19922264220205

Acknowledgments

This work was supported in part by RFBR grants 19-07-00356, 18-07-00944, and 20-07-00537.

References

1. Torshin, I. Yu. 2022. O primenenii topologicheskogo podkhoda k analizu plokho formalizuemykh zadach dlya postroeniya algoritmov virtual'nogo skrininga kvantovo-mekhanicheskikh svoystv organicheskikh molekul I: Osnovy problemno orientirovannoy teorii [On the application of a topological approach to analysis of poorly formalized problems for constructing algorithms for virtual screening of quantum-mechanical properties of organic molecules I: The basics of the problem-oriented theory]. *Informatika i ee Primeneniya — Inform Appl.* 15(1):39–45.

2. Stepanov, N. F. 2001. *Kvantovaya mekhanika i kvantovaya khimiya* [Quantum mechanics and quantum chemistry]. Moscow: Mir. 519 p.

3. Torshin, I. Yu., and K. V. Rudakov. 2014. On the application of the combinatorial theory of solvability to the analysis of chemographs. Part 1: Fundamentals of modern chemical bonding theory and the concept of the chemograph. *Pattern Recognition Image Analysis* 24(1):11–23.
4. Torshin, I. Yu., and K. V. Rudakov. 2019. On the procedures of generation of numerical features over partitions of sets of objects in the problem of predicting numerical target variables. *Pattern Recognition Image Analysis* 29(4):654–667. doi: 10.1134/S1054661819040175.
5. Ramakrishnan, R., P. Dral, and M. Rupp. 2014. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* 1(1):140022. 7 p. doi: 10.1038/sdata.2014.22.
6. Torshin, I. Yu., and K. V. Rudakov. 2014. On the application of the combinatorial theory of solvability to the analysis of chemographs. Part 2: Local completeness of invariants of chemographs in view of the combinatorial theory of solvability. *Pattern Recognition Image Analysis* 24(2):196–208.

Received April 5, 2021

Contributor

Torshin Ivan Y. (b. 1972) — Candidate of Science (PhD) in physics and mathematics, Candidate of Science (PhD) in chemistry, senior scientist, A. A. Dorodnicyn Computing Center, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 40 Vavilov Str., Moscow 119333, Russian Federation; tiy135@yahoo.com