ЖУРНАЛ ОБЩЕЙ ХИМИИ, 2021, том 91, № 3, с. 449–473

УДК 544.35;544.341.2;519.254

КОМПЬЮТЕРНЫЙ АНАЛИЗ УСТОЙЧИВОСТИ КРИПТАТОВ $M[222]^+$ КАТИОНОВ ЩЕЛОЧНЫХ МЕТАЛЛОВ В РАЗНЫХ РАСТВОРИТЕЛЯХ

© 2021 г. Н. В. Бондарев*

Харьковский национальный университет имени В. Н. Каразина, пл. Свободы 4, Харьков, 61022 Украина
**e-mail: bondarev_n@rambler.ru*

Поступило в Редакцию 22 декабря 2020 г.
После доработки 22 декабря 2020 г.
Принято к печати 15 января 2021 г.

Проведен компьютерный анализ термодинамических констант комплексообразования криптанда [222] с катионами щелочных металлов (криптаты $M[222]^+$, где M = Li, Na, K, Rb, Cs) в воде и органических растворителях – метаноле, этаноле, 1-пропаноле, ацетонитриле, бензонитриле, ацетоне, N,N-диметилформамиде, N-метилпирролидоне, нитробензоле, нитрометане, 1,2-дихлорэтане, пропиленкарбонате при 298.15 K. Построены разведочные (факторная, кластерные, дискриминантная, каноническая, дерево решений), регрессионные и нейросетевые модели влияния свойств растворителей и катионов на устойчивость криптатов катионов. Обучены нейросетевые аппроксиматор MLP 4-7-1 и классификаторы констант устойчивости криптатов – многослойный персептрон MLP 4-7-4 и самоорганизующаяся сеть Кохонена SOFM 8-4. На независимых данных по константам устойчивости криптатов катионов щелочных металлов демонстрируются прогностические возможности обученного персептрона-аппроксиматора MLP 4-7-1.

Ключевые слова: криптанд [222], константа комплексообразования, разведочный анализ, множественная линейная регрессия, нейронные сети, моделирование, прогнозирование

DOI: 10.31857/S0044460X21030112

Открытие синтетических макроциклических соединений, таких как краун-эфиры, криптанды, сферанды [1–3], способных образовывать комплексы типа хозяин–гость с ионами металлов, анионами и органическими молекулами, положило начало супрамолекулярной химии [4, 5]. Связывание химических форм за счет нековалентного взаимодействия лежит в основе образования супрамолекул, которые характеризуются определенной термодинамической устойчивостью.

Несмотря на большой объем данных по константам устойчивости комплексов макроциклических лигандов с катионами металлов в растворах, остаются актуальными следующие вопросы термодинамики супрамолекулярных комплексов [6, 7]: прогнозирование констант устойчивости комплексов краун-эфиров и криптандов с катионами в зависимости от строения лиганда, что необходимо для практического конструирования лигандов с заданной селективностью комплексообразования и определенной устойчивостью их комплексов; оценка и прогнозирование устойчивости комплексов при замене растворителя, что требуется для практических целей разделения химических форм; оценка и прогнозирование констант устойчивости при переходе от одного катиона к другому, поскольку такие расчеты методами молекулярной динамики трудоемки.

В продолжение предыдущих работ [8–13] здесь сообщаются результаты компьютерного анализа термодинамических констант комплексообразования криптанда [222] (рис. 1) с катионами щелочных металлов (криптаты $M[222]^+$, где M = Li, Na, K, Rb, Cs) в неводных растворителях с целью

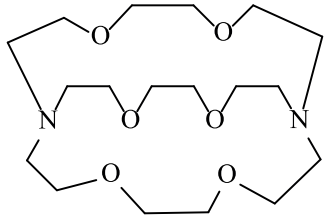

Рис. 1. Криптанд [222] – (4,7,13,16,21,24-гексаокса-1,2,10-диазабицикло[8.8.8]гексакозан.

построения разведочных, регрессионных и нейросетевых моделей влияния свойств растворителей и катионов на устойчивость криптатов катионов и прогнозирования констант комплексообразования в еще не исследованных средах.

Для проведения компьютерного моделирования использованы литературные данные по константам устойчивости комплексов $M[222]^+$ из обзора И. Маркуса [14].

Мерой устойчивости комплексов криптанда [222] (L) с катионами является термодинамическая константа устойчивости $K = [LM^+]\gamma_{LM+}/[L]\gamma_L[M^+]\gamma_{M+}$, отвечающая простейшей схеме комплексообразования: $L_s + M_s^+ = LM_s^+$, где [L] и γ_L, $[M^+]$ и γ_{M+}, $[LM^+]$ и γ_{LM+} – равновесные концентрации и коэффициенты активности свободного лиганда (L), катиона и комплекса соответственно; s – растворитель. Заряд ионов, участвующих в равновесии комплексообразования, одинаков и не изменяется (+1), концентрационная константа устойчивости практически не зависит от ионной силы раствора при концентрациях < 0.01 моль·дм$^{-3}$, поэтому приравнивается термодинамической. Предполагается, что катионы M_s^+ и $M[222]_s^+$ не образуют ионные ассоциаты с анионами больших размеров(BF_4^-, ClO_4^- или $CF3SO_3^-$).

Значения констант устойчивости криптатов катионов, полученные разными авторами, обычно были в пределах ± 0.2 лог. ед., для $3 < \lg K < 10$. Автором [14], отмечены выбросы из общей выборки констант устойчивости криптатов: заниженные значения для $\lg K$ $Cs[222]^+$ в диметилсульфоксиде, завышенные значения $\lg K$ для $Na[222]^+$ и $Cs[222]^+$ и заниженные значения $\lg K$ для $K[222]^+$ комплексов в ацетонитриле.

В криптанде [222] два атома азота третичного амина соединены тремя цепочками –$C_2H_4OC_2H_4OC_2H_4$–, образующими клетку с десольватированным катионом металла, удерживаемым восьмью электронодонорными атомами – шестью атомами кислорода и двумя атомами азота. Известно, что криптанд [222] существует в трех конформациях: *экзо-экзо*, *экзо-эндо* и *эндо-эндо*, в зависимости от положения неподеленных электронных пар на третичных атомах азота, вне (*экзо*) или внутри (*эндо*) полости [14]. Поэтому образование комплексов катионов с криптандом [222] сопровождается конформационной предорганизацией последнего.

Для описания свойств растворителей использованы физические константы: ε – относительная диэлектрическая проницаемость [15]; $\Delta_v H$ – энтальпия испарения, кДж/моль [15]; δ^2– плотность энергии когезии [15], Дж/см^3; D_S – диаметр молекулы растворителя, нм [15]; V_{in} – внутренний объем одного моля растворителя, см^3/моль [15]. Для количественной оценки донорной и акцепторной эффективности растворителей использованы спектроскопические эмпирические параметры полярности Камлета–Тафта β [15] и Димрота–Райхардта E_T [15]. Физические свойства катионов представлены в работе радиусом катиона r_M, нм [16] и энергией ионизации атомов щелочных металлов U_1, эВ [17, 18].

Компьютерное моделирование констант устойчивости криптатов $M[222]^+$ проведено в средах STATISTICA 12 и SPSS 23 на платформе Windows 10 для комплексов состава 1:1 криптанда [222] с катионами щелочных металлов в воде (W) и органических растворителях (метаноле, этаноле, 1-пропаноле, ацетонитриле, бензонитриле, ацетоне, N,N-диметилформамиде, N-метилпирролидоне, нитробензоле, нитрометане, 1,2-дихлорэтане, пропиленкарбонате) при 298.15 K.

Поставленная цель достигнута путем решения следующих задач: (1) первичный анализ данных, вычисление описательных статистик, проверка нормальности распределения; (2) факторный анализ – построение корреляционных матриц, выделение латентных факторов; (3) кластерный анализ – алгоритм древовидной кластеризации, итерационный алгоритм *k*-средних; (4) дискриминантный анализ Фишера – построение линейных классификационных функций; (5) канонический дискриминантный анализ – построение канони-

Таблица 1. Описательная статистика показателей комплексообразования, отобранных для разведочного анализа

Показатель	Количество значений	Среднее	Минимальное значение	Максимальное значение	Стандартное отклонение	Стандартная ошибка
lg$K$	64	7.3	1.0	13.6	3.4	0.42
r_M	5	0.1	0.1	0.2	0.0	0.00
ε	13	35.7	10.4	78.4	18.1	2.27
D_S	13	0.5	0.3	0.6	0.1	0.01
V_{in}	13	54.6	16.7	87.1	19.2	2.40
β	13	0.5	0.1	0.9	0.3	0.03
E_T	13	0.5	0.3	1.0	0.2	0.02
δ^2	13	724.5	400.0	2294.4	474.6	59.33
$\Delta_v H$	13	44.8	31.0	65.3	9.7	1.22
U_1	5	4.6	3.9	5.4	0.6	0.07

ческих линейных дискриминантных функций; (6) деревья классификации – построение дендрограммы и правила кластеризации устойчивости криптатов; (7) регрессионный анализ зависимости устойчивости криптатов от свойств растворителей и катионов; (8) нейросетевой анализ – нейросетевые классификатор, нейросетевой аппроксиматор; (9) аппроксимирующие и прогностические возможности регрессионных и нейросетевых моделей.

Первичный анализ данных. В табл. 1 приведены количественные параметры описательной статистики [19] отобранных для анализа показателей. Распределение данных можно считать симметричным, если среднее квадратическое отклонение (стандартное отклонение) данных меньше половины среднего арифметического. Проверка гипотезы нормального распределения анализируемых данных (табл. 2) выполнена по критериям Шапиро–Уилка ($8 < n < 50$) и Колмогорова–Смирнова ($n > 50$) [20].

В основу компьютерного анализа данных положены математические приемы, статистические методы и практические рекомендации, изложенные в [21–24].

Факторный анализ. Задача анализа состояла в изучении структуры взаимосвязей признаков (свойств системы в метрической шкале), уменьшении их исходного количества путем перехода к новым переменным – факторам и отборе ортогональных дескрипторов. Фактор при этом интерпретируется как причина совместной изменчивости нескольких исходных переменных. В один фактор объединяются переменные, сильно коррелирующие между собой. Поэтому основное назначение факторного анализа – анализ корреляций множества переменных, не разделяемых на независимые и зависимые.

В ходе проведения факторного анализа были рассчитаны: а) критерий КМО – мера адекватности выборки Кайзера–Мейера–Олкина [19] – показатель, используемый для оценки применимости факторного анализа. Значения от 0.5 до 1 свидетельствуют об адекватности факторного анализа, значения меньше 0.5 указывают на то, что факторный анализ неприменим к выборке; б) критерий сферичности Бартлетта [19] – показатель, который позволяет проверить, отличаются ли корреляции от 0. Если коэффициент корреляции близок к

Таблица 2. Расчетные и табличные (критические) значения критериев проверки гипотезы нормальности распределения переменных[а]

Переменная, (n)	Критерий Шапиро–Уилка, $W_{расч}$ ($W_{табл}$)
lg$K$ (64)	б
ε (13)	0.868 (0.874)
D_S (13)	0.939 (0.874)
V_{in} (13)	0.955 (0.874)
β (13)	0.910 (0.874)
E_T (13)	0.824 (0.874)
δ^2 (13)	0.517 (0.874)
$\Delta_v H$ (13)	0.933 (0.874)

[а] n – объем выборки, p – уровень значимости. Если табличное значение $W_{табл}$ меньше расчетного значения $W_{расч}$, а $D_{табл} > D_{расч}$, то распределение считается соответствующим нормальному на уровне значимости p 0.05.

[б] Критерий Колмогорова–Смирнова, $D_{расч}$ ($D_{табл}$) 0.091 (0.166).

Таблица 3. Корреляционная матрица переменных

Переменная	lg$K$	r_M	ε	D_S	V_{in}	β	E_T	δ^2	$\Delta_v H$	U_1
lg$K$	1.00	0.03	–0.36	0.22	0.23	–0.46	–0.35	–0.41	–0.11	–0.02
r_M	0.03	1.00	0.00	0.01	0.00	0.02	–0.02	–0.01	0.01	–0.99
ε	–0.36	0.00	1.00	–0.39	–0.38	0.00	0.55	0.70	0.44	–0.00
D_S	0.22	0.01	–0.39	1.00	0.99	0.13	–0.79	–0.74	0.59	–0.01
V_{in}	0.23	0.00	–0.38	0.99	1.00	0.06	–0.80	–0.68	0.57	–0.00
β	–0.46	0.02	0.00	0.13	0.06	1.00	0.24	0.05	0.29	–0.01
E_T	–0.35	–0.02	0.55	–0.79	–0.80	0.24	1.00	0.84	–0.10	0.01
δ^2	–0.41	–0.01	0.70	–0.74	–0.68	0.05	0.84	1.00	–0.06	0.01
$\Delta_v H$	–0.11	0.01	0.44	0.59	0.57	0.29	–0.10	–0.06	1.00	–0.01
U_1	–0.02	–0.99	–0.00	–0.01	–0.00	–0.01	0.01	0.01	–0.01	1.00

Таблица 4. Матрица компонентных нагрузок (факторная структура) до вращения факторов

Переменная (свойство)	Факторные нагрузки до вращения (a_{ik}) Извлечение факторов методом главных компонент			
	Фактор 1	Фактор 2	Фактор 3	Фактор 4
lg$K$	–0.440	0.225	–0.489	–0.437
r_M	–0.023	–0.889	–0.452	0.002
ε	0.619	–0.251	0.455	–0.528
D_S	–0.927	–0.156	0.327	–0.036
V_{in}	–0.915	–0.137	0.306	–0.082
β	0.067	–0.297	0.541	0.689
E_T	0.919	–0.068	0.119	0.041
δ^2	0.900	–0.096	0.159	–0.183
$\Delta_v H$	–0.289	–0.394	0.769	–0.367
U_1	0.022	0.889	0.452	–0.004
Собственное значение[a]	4.018	1.993	1.979	1.122
Доля дисперсии	0.402	0.199	0.198	0.112

[a] Сумма квадратов факторных нагрузок (a_{ik}^2).

нулю, то выбранная переменная не взаимосвязана с другими. Уровень значимости p меньше 0.05 указывает, на то что проведение факторного анализа приемлемо; в) корреляционная матрица [19] – матрица, содержащая все возможные коэффициенты парных корреляций между анализируемыми переменными (свойствами) табл. 3. Рассчитанная мера выборочной адекватности Кайзера–Мейера–Олкина равна 0.520; значения критерия сферичности Бартлетта: Хи-квадрат (приближенный) – 977.0 для числа степеней свободы 45, уровень значимости, p 0.000. Величина КМО показывает приемлемую адекватность выборки для факторного анализа КМО 0.520 > 0.5. Критерий Бартлетта $p < 0.05$, что свидетельствует о целесообразности факторного анализа в силу наличия достаточной коррелированности переменных [24].

Матрица интеркорреляций исходных данных (табл. 3) обработана с использованием анализа главных компонент [21, 22]. Основной принцип выделения латентных факторов методом главных компонент – представление двух или более зависимых переменных одним фактором. В основе анализа главных компонент лежит математический метод нахождения собственных значений и собственных векторов корреляционной матрицы. Собственные значения λ – дисперсии (изменчивости), выделяемые факторами. Название связано с алгебраическим способом вычисления λ при решении матрично-векторного уравнения AV = λV [25], где A – линейный оператор в матричной форме (матрица корреляций), V – собственный вектор линейного преобразования A, λ – собственное значение (число) линейного оператора A, λV – колли-

Таблица 5. Общности переменны a_{ik}^2 (свойств системы)

Номер переменной	Переменная	Фактор 1	Фактор 2	Фактор 3	Фактор 4	R^2 [a]
1	lg$K$	0.194	0.245	0.484	0.675	0.423
2	r_M	0.001	0.790	0.995	0.995	0.979
3	ε	0.384	0.446	0.654	0.933	0.969
4	D_S	0.858	0.883	0.990	0.991	0.996
5	V_{in}	0.838	0.857	0.950	0.957	0.988
6	β	0.005	0.092	0.385	0.860	0.617
7	E_T	0.844	0.849	0.863	0.865	0.977
8	δ^2	0.811	0.820	0.845	0.879	0.917
9	$\Delta_v H$	0.084	0.239	0.829	0.964	0.986
10	U_1	0.000	0.790	0.995	0.995	0.979

[a] R^2 – коэффициент множественной детерминации.

неарный вектор. Результатом решения уравнения является матрица компонентных нагрузок (табл. 4).

Различия в методах факторного анализа определяются тем, как решается проблема общностей. Единичная дисперсия каждой переменной представлена в факторном анализе как сумма ее общности и характерности [21, 22].

$$1 = h_i^2 + e_i^2.$$

Здесь h^2 – общность переменной с номером i (от 1 до 10) для фактора k (от 1 до 4) табл. 5; e^2 – ее характерность.

Общность – это часть дисперсии переменной, обусловленная действием общих факторов, иначе говоря общность является квадратом множественной корреляции переменной как зависимой, использующей факторы как предикторы.

Общность переменной i равна сумме квадратов ее нагрузок (табл. 4) по общим факторам (по строке факторных нагрузок) [21, 22]:

$$h_i^2 = \sum_{k=1}^{4} a_{ik}^2.$$

Характерность – часть ее дисперсии, обусловленная спецификой данной переменной и ошибками измерения (разность полной единичной дисперсии переменной и ее общности).

Любой элемент факторной структуры – факторная нагрузка переменной (табл. 4), возведенная в квадрат (табл. 5) – приобретает смысл доли дисперсии переменной, обусловленной данным фактором [19]. Суммирование этих долей по строке дает общность – долю дисперсии (изменчивости) переменной, обусловленную влиянием четырех общих факторов. Суммирование долей дисперсии всех переменных по одному фактору дает суммарную дисперсию всех переменных, обусловленную действием этого фактора, что равно количеству переменных (10).

Факторная структура до вращения не интерпретируется, однако содержит важную информацию – суммарную долю дисперсии (информативность) факторов и значения общностей переменных (свойств системы). Суммарная доля дисперсии – показатель того, насколько полно выделяемые факторы могут представить данный набор свойств системы и наоборот, набор свойств – выделяемые факторы. Общность переменной (строки табл. 5) – показатель влияния переменной (свойства) на факторную структуру [21, 22].

Сумма квадратов всех элементов факторной структуры (факторных нагрузок) – равна сумме всех общностей и суммарной дисперсии всех переменных, обусловленной общими факторами. Эта величина, деленная на количество переменных, известна как полнота факторизации V [21, 22]:

$$V = \sum_{k=1}^{M} V_k = \frac{1}{P}\sum_{k=1}^{M} \lambda_k = \frac{1}{P}\sum_{k=1}^{M} h_i^2,$$

где V_k – мощность фактора с номером k; λ_k – собственное число фактора с номером k; $h^2{}_i$ – общность переменной i; M = 4 – число факторов; P = 10 – число переменных (свойств).

Для упрощения интерпретации выделенных факторов в работе использован ортогональный ме-

Таблица 6. Объясненная совокупная дисперсия (изменчивость)

Компонент	Начальные собственные значения выделенных факторов до вращения)			Начальные собственные значения главных факторов до вращения			Собственные значения главных факторов после вращения		
	Собственное значение, λ	% объясненной дисперсии (изменчивости)	% совокупного собственного значения	Собственное значение, λ	% объясненной дисперсии (изменчивости)	% совокупного собственного значения	Собственное значение, λ	% объясненной дисперсии (изменчивости)	% совокупного собственного значения
1	4.018	40.184	40.184	4.018	40.184	40.184	3.944	39.442	39.442
2	1.993	19.928	60.113	1.993	19.928	60.113	1.991	19.906	59.348
3	1.979	19.789	79.902	1.979	19.789	79.902	1.640	16.401	75.749
4	1.122	11.216	91.118	1.122	11.216	91.118	1.537	15.369	91.118
5	0.545	5.453	96.570						
6	0.200	2.001	98.571						
7	0.120	1.205	99.776						
8	0.011	0.105	99.882						
9	0.009	0.091	99.973						
10	0.003	0.027	100.0						

тод вращения (метод варимакс) [21, 22]), минимизирующий число переменных с высокими нагрузками на каждый фактор. Такое вращение факторов ведет к максимизации дисперсии (изменчивости) «новой» переменной (фактора) и минимизации разброса переменных вокруг нее. Факторы последовательно выделяются один за другим. Каждый последующий фактор определяется так, чтобы максимизировать изменчивость (варимакс), оставшуюся от выделения предыдущих факторов. Поэтому факторы оказываются независимыми друг от друга – некоррелированными (ортогональными).

При повторных итерациях выделяются факторы с все меньшей и меньшей дисперсией (табл. 4). Итерационная процедура начиналась с матрицы, в которой дисперсия (изменчивость) каждой переменной равна 1. Поэтому общая дисперсия равна числу переменных – 10, т. е. наибольшей изменчивости, которая может быть выделена. Изменчивость, объясненная последовательно выделяемыми факторами до и после вращения, представлена в табл. 6.

Наличие информации о том, сколько дисперсии (изменчивости) выделил каждый фактор (табл. 4) позволило оставить 4 фактора (четыре главные компоненты) на основе критерия Х. Кайзера [26] и графического метода Р. Кэттелля [27] – критерия каменистой осыпи. По критерию Кайзера отбираются факторы с собственными значениями больше 1. Согласно Кэттеллю, на графике зависимости собственных значений от числа факторов находится точка, где убывание собственных значений слева направо замедляется.

Качество факторного анализа тем выше, чем выше полнота факторизации. Построенная факторная модель сохраняет 91.12% (табл. 4, 6) исходной информации, при этом число факторов сокращается в два с половиной раза с 10 до 4.

Результат варимакс-вращения главных факторов представлен в табл. 7. Как отмечено ранее, строки таблицы содержат факторные нагрузки переменных (10 свойств изучаемой системы) по четырем факторам (столбцам). Факторные нагрузки в этом случае, являющиеся аналогом коэффициента корреляции, изменяются от –1 до +1 и показывают степень взаимосвязи соответствующих переменных и факторов – чем больше абсолютная величина факторной нагрузки, тем сильнее связь

Таблица 7. Факторные нагрузки после варимакс-вращения

Переменная (свойство)	Факторные нагрузки (a_{ik})			
	Фактор 1	Фактор 2	Фактор 3	Фактор 4
lg$K$	0.306	0.029	–0.210	–0.732
r_M	0.005	0.997	0.002	0.000
ε	–0.477	0.002	0.840	–0.001
D_S	0.986	0.007	0.123	0.064
V_{in}	0.969	–0.001	0.133	0.014
β	0.073	0.018	0.040	0.923
E_T	–0.862	–0.010	0.248	0.246
δ^2	–0.831	–0.004	0.422	0.100
$\Delta_v H$	0.486	0.004	0.827	0.207
U_1	–0.004	–0.997	–0.001	–0.002
Собственноезначение, λ	3.906	1.991	1.708	1.507
Доля дисперсии	0.391	0.199	0.171	0.151

переменной с фактором, тем больше данная переменная обусловлена действием соответствующего фактора [21, 22].

Поскольку по Фактору 1 максимальные нагрузки имеют переменные D_S, V_{in}, E_T, δ^2, то Фактору 1 (новой переменной) может быть присвоено название «свойства растворителя 1». Фактору 2 можно присвоить название «свойства катиона» (r_M и U_1). Аналогично Фактор 3 можно назвать «свойства растворителя 2» – ε и Δ_vH. Фактор 4 коррелирует со свойствами равновесия комплексообразования и растворителя – lg$K$ и β. Нетрудно заметить, что переменные, определяющие фактор, сильнее связаны друг с другом, чем с другими переменными (табл. 3). Таким образом, за взаимосвязью десяти исходных показателей исследуемого процесса образования криптатов катионов в разных средах при помощи факторного анализа обнаруживается действие четырех новых латентных переменных (факторов), объединяющих переменные (свойства системы) в группы по степени влияния на факторы.

На заключительном этапе факторного анализа, вместо вычисления значений факторов, выбраны переменные-заменители (surrogate variables), характеризуемые большими нагрузками (высокими коэффициентами корреляции) на каждый ортогональный фактор, для использования их в последующем анализе.

Перебор возможных комбинаций ортогональных дескрипторов показал, что оптимальными и наиболее информативными с точки зрения химической природы (химизма) исследуемой зависимости устойчивости криптатов M[222]$^+$ от свойств среды (растворителей различной химической природы) и реагентов (катионов щелочных металлов) являются эмпирические параметры: из Фактора 1 – E_T (–0.862), из Фактора 2 – U_1 (–0.997), из Фактора 3 – ε (0.840), из Фактора 4 – β (0.923) (табл. 7).

Результаты факторного анализа были положены в основу проведения кластерного, дискриминантного, канонического, построение дерева классификации, регрессионного и нейросетевого анализа влияния свойств растворителя и катиона на константу комплексообразования криптанда [222] с катионами щелочных металлов.

Кластерный анализ. На языке математики задача кластерного анализа заключается в том, чтобы на основании данных, содержащихся во множестве X, разбить множество объектов $X_i \in X$ на m (m – целое) кластеров (подмножеств) Q_1, Q_2, ..., Q_m так, чтобы каждый объект X_j принадлежал одному и только одному подмножеству разбиения (кластеру Q_l) и чтобы объекты, принадлежащие одному и тому же кластеру, были сходными, а объекты, принадлежащие разным кластерам – разнородными [21, 22].

Кластер имеет следующие математические характеристики [21, 22]: центр, радиус, среднеквадратическое отклонение, размер кластера. Центр кластера – это среднее геометрическое место то-

чек в пространстве переменных. Радиус кластера – максимальное расстояние точек от центра кластера. Кластеры могут быть перекрывающимися, если они содержат спорные объекты. Размер кластера определяется либо по радиусу кластера, либо по среднеквадратичному отклонению объектов для этого кластера. Объект относится к кластеру, если расстояние от объекта до центра кластера меньше радиуса кластера. Если это условие выполняется для двух и более кластеров, объект является спорным.

Применительно к рассматриваемому равновесию комплексообразования, множество X – это перечень из n = 64 наблюдений констант устойчивости криптатов катионов $M[222]^+$ (объектов K_1, K_2, K_3,…, K_n) в разных растворителях, выбранных для кластерного анализа. Каждое из этих наблюдений охарактеризовано набором химических показателей (5 свойств –признаков), выраженных в числовой форме: величина константы устойчивости и связанные с ней свойства растворителей – ε, E_T, β и катионов – U_1. Тогда X_1 (вектор измерений) представляет собой набор указанных характеристик для первого наблюдения: $X_1 = (x_{1,1}, x_{1,2}, x_{1,3}, x_{1,4}, x_{1,5})$, X_2 – для второго, X_3 – для третьего, и т. д.

Решением задачи кластерного анализа является разбиение множества элементов матрицы признаков размером 64×5 на m групп (m кластеров) – подмножеств, удовлетворяющих критерию оптимальности:

– каждая константа комплексообразования (объект) должна принадлежать одному и только одному подмножеству разбиения (кластеру);

– константы комплексообразования, принадлежащие одному и тому же кластеру, должны быть сходными – количественно характеризовать равновесие комплексообразования $M_s^+ + [222]_s = M[222]_s^+$ в одних и тех же растворителях (растворителе);

– константы устойчивости криптатов, принадлежащие разным кластерам, должны быть разнородными – содержать разные наборы как констант устойчивости криптатов, так и растворителей.

Константы устойчивости криптатов, подлежащие кластеризации, представляются точками в p-мерном пространстве признаков. Тогда сходство между объектами определяется через понятие расстояния между точками, чем меньше расстояние между объектами, тем они более схожи [21, 22].

В качестве меры сходства констант устойчивости криптатов использована евклидова метрика [22, 23]. Евклидово расстояние $d(X_i, X_j)$ – геометрическое расстояние между парами векторов X_i и X_j (константами устойчивости) в p-мерном пространстве признаков (свойств растворителей и катионов (x_{ki} и x_{kj}), по которым сравниваются константы устойчивости. Евклидово расстояние равно квадратному корню из суммы квадратов разностей значений для каждой переменной (свойства) [19]:

$$d(X_i, X_j) = \left[\sum_{k=1}^{n}(x_{ki} - x_{kj})^2\right]^{\frac{1}{2}},$$

Перед проведением кластерного анализа констант устойчивости комплексов $M[222]^+$ была выполнена стандартизация данных [19, 23]. При этом шкала измерения каждой переменной изменяется таким образом, чтобы среднее равнялось нулю, а стандартное отклонение – единице (z преобразование):

$$z = \left(x - \bar{x}\right) / \sigma,$$

где $\bar{x}$ и σ – среднее и среднеквадратическое отклонение переменной x соответственно.

Кластеризация констант устойчивости криптатов по свойствам растворителей и катионов проведена двумя методами: агломеративным – метод Варда и итеративным – метод k-средних [21].

Агломеративная кластеризация начинается с размещения каждой константы устойчивости (объекта) в отдельном кластере. Затем кластеры объединяются, группируя константы устойчивости каждый раз во все более и более крупные кластеры. Этот процесс продолжается до тех пор, пока все константы устойчивости (объекты) не станут набором одного единственного кластера рис. 2. Наблюдаемые (экспериментальные) константы устойчивости обозначены как C_i, где i – номер константы комплексообразования.

В качестве целевой функции, представляющей собой функционал, выражающий уровни желательности различных разбиений и группировок, в методе Варда применяется внутригрупповая сумма квадратов отклонений, вычисляемая как сумма квадратов расстояний между каждой точкой (константой устойчивости криптата) и средней по кластеру, содержащему эту константу [21]:

$$W = \sum_{j=1}^{n}(x_j - \overline{x})^2 = \sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2,$$

где x_j – свойства растворителей и катионов, связанные с j-ой константой устойчивости криптата.

В качестве расстояния $dis(X,Y)$ между кластерами X и Y берется прирост суммы квадратов расстояний объектов (констант устойчивости) до центров кластеров, получаемый в результате их объединения [21, 22]:

$$dis(X,Y) = \frac{n_x n_y}{n_x + n_y}(\overline{X} + \overline{Y})^T(\overline{X} + \overline{Y}),$$

где X, Y – радиусы-векторы центров кластеров, n_x, n_y – число элементов в них, верхний индекс T означает транспонирование. Метод Варда [21, 22] минимизирует сумму квадратов отклонений для любых двух (гипотетических) кластеров, которые могут быть сформированы. На каждом шаге объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т. е. внутригрупповой суммы квадратов отклонений.

Дендрограмма (рис. 2) показывает, что в результате кластеризации константы устойчивости коронатов $M[222]^+$ группируются в два, три или четыре явно выраженных кластера в зависимости от расстояния объединения.

Важно отметить, если число кластеров определено, то можно получить прогнозную информацию о принадлежности константы устойчивости к определенному классу (группе) по свойствам растворителей и катионов.

Метод *k*-средних. Начальные разбиения на кластеры, требующие детального распределения данных о константах устойчивости по группам, задавались не случайным образом, а на основе решения, полученного иерархической кластеризацией [21] методом Варда – 2, 3 или 4 кластера (рис. 2). Итерации по принципу *k*-средних начинались последовательно с двух, трех или четырех выбранных кластеров, а затем изменялась принадлежность объектов к ним, чтобы, во-первых – минимизировать изменчивость внутри кластеров и, во-вторых – максимизировать изменчивость между кластерами. Мерой изменчивости выступает сумма квадратов (Sum Squeres) отклонений от среднего:

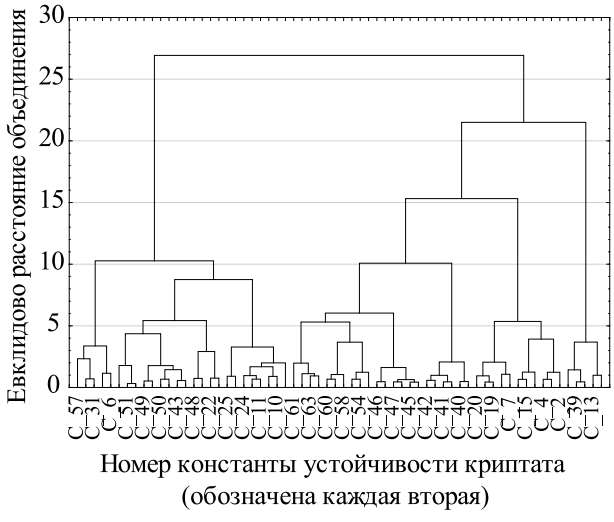

Рис. 2. Дендрограмма кластеризации констант устойчивости криптатов катионов по методу Варда в пакете STATISTICA 12.

$$SS = \sum(x_i - \overline{x})^2.$$

Анализ результатов метода *k*-средних показал, что в наибольшей мере критерию оптимальности отвечает разбиение множества X на четыре кластера: 1 кластер – содержит 29 констант устойчивости криптатов $M[222]^+$ (M = Li, Na, K, Rb, Cs) в протолитических (MeOH, EtOH, PrOH) и апротонных (N-метилпирролидон, ДМСО, ДМФА) растворителях; 2 кластер – содержит 5 констант устойчивости коронатов $M[222]^+$ (M = Li, Na, K, Rb, Cs) в воде; 3 кластер – содержит 12 констант устойчивости коронатов $Li[222]^+$ и $Na[222]^+$ в апротонных растворителях: ацетоне, пропиленкарбонате, нитрометане, ацетонитриле, бензонитриле, 1,2-дихлорэтане; 4 кластер – содержит 18 констант устойчивости коронатов $K[222]^+$, $Rb[222]^+$ и $Cs[222]^+$ в тех же апротонных растворителях: ацетоне, пропиленкарбонате, нитрометане, ацетонитриле, бензонитриле, 1,2-дихлорэтане.

В табл. 8 приведены значения межгрупповых SS_B и внутригрупповых SS_W сумм квадратов отклонений от среднего для каждой переменной. Чем меньше SS_W и больше значение SS_B, тем лучше переменная характеризует принадлежность объектов к кластеру и тем качественнее кластеризация. SS_W – сумма квадратов отклонений значений каждого из предикторов (свойства растворителей и катионов) от группового среднего значения предиктора внутри группы (кластера) – мера внутригрупповой из-

Таблица 8. Результаты дисперсионного анализа стандартизированных показателей комплексообразования методом *k*-средних

Переменная	SS_B	df_B	SS_W	df_W	$F(3, 60)$	p
lg$K$	20.51	3	42.49	60	9.65	0.000
ε	29.98	3	33.02	60	18.15	0.000
β	48.22	3	14.78	60	65.23	0.000
E_T	41.86	3	21.14	60	39.62	0.000
U_1	27.61	3	35.39	60	15.60	0.000

Таблица 9. Центры четырех кластеров констант устойчивости криптатов катионов[a]

Кластер	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cluster 1	0.00	2.84	1.23	1.02
Cluster 2	1.69	0.00	4.22	4.03
Cluster 3	1.11	2.05	0.00	0.77
Cluster 4	1.01	2.01	0.88	0.00

[a] Евклидовы расстояния (ниже диагонали) и квадраты евклидовых расстояний (выше диагонали).

менчивости:

$$\sigma^2_{SSw} = SS_W / (n-1),$$

где σ^2_{SSW} – внутригрупповая дисперсия; SS_B – межгрупповая сумма квадратов отклонений средних значений предикторов в каждой из групп от суммарного среднего значения предикторов по всем группам – мера межгрупповой изменчивости: $\sigma^2_{SSB} = SS_B/(n-1)$, где σ^2_{SSB} – межгрупповая дисперсия; значение критерия Фишера $F = MS_B/MS_W$, где $MS_B = SS_B/df_B$, $MS_W = SS_W/df_W$; MS_W и MS_B – средние значения квадратов отклонений внутри групп и между ними; $df_W = (n - m - 1)$ и $df_B = (m - 1)$ – соответствующие степени свободы (m – число групп, n – количество наблюдений в каждой из групп); $F(3, 60)$ – наблюдаемый критерий Фишера. [$F_{кр}(3, 60, p\ 0.05) = 2.76$]; p – наблюдаемый уровень значимости [21–23].

По статистическим показателям, полученным в результате дисперсионного анализа (табл. 8), кластеры являются различимыми, так как для всех переменных $F_{набл}(3, 60) > F_{кр}(3, 60)$, а $p < 0$. Однако близкие значения центров второго и четвертого кластеров (табл. 9), а также значения межгрупповых SS_B, которые меньше внутригрупповых SS_W для свойств ε и U_1 (табл. 8), свидетельствуют о том, что для подтверждения результатов методов Варда и *k*-средних, необходимо другими методами предоставить убедительные доказательства правильности разделения констант устойчивости на четыре группы.

С этой целью проведены дискриминантный, канонический, нейросетевой анализы и построено дерево классификации.

Множественный дискриминантный анализ. Дискриминантный анализ используется для изучения различий между несколькими группами по определенному набору дискриминантных переменных. Математические дефиниции и допущения дискриминантного анализа [21]: g – число классов (кластеров); p – число дискриминантных переменных; n_i – число объектов (наблюдений) класса i; n – общее число объектов всех классов. В модели дискриминантного анализа должно быть: (а) два или более классов – $g \geq 2$; (б) по крайней мере наличие двух объекта в каждом классе – $n_i \geq 2$; (в) любое число дискриминантных переменных при условии, что оно не превосходит общее число объектов за вычетом двух – $0 < p < (n - 2)$; (г) измерение дискриминантных переменных по интервальной шкале; (д) линейная независимость дискриминантных переменных; (е) приблизительное равенство ковариационных матриц для каждого класса; (ж) многомерная нормальность закона распределения.

Математическая постановка задачи дискриминантного анализа состояла в следующем [21, 22]. Имеется n объектов с m характеристиками. В результате измерений каждый объект характеризует-

Таблица 10. Достоверность различения четырех групп констант устойчивости криптатов M[222]$^+$ по каждой переменной

Свойство	Группирующая переменная: 4 кластера констант устойчивости криптатов; Λ-Уилкса 0.035; приближенное значение $F_{набл}$(28, 151) 32.119, p < 0.000; $F_{кр}$(28, 151) 1.8, $F_{кр}$(3, 57) 2.76					
	Λ Уилкса	частная Λ Уилкса	$F_{исключить}$(3, 57)	p-уровень	толерантность,1-R^2	R^2
ε	0.047	0.750	6.335	0.001	0.970	0.030
β	0.140	0.249	57.157	0.000	0.977	0.023
E_T	0.068	0.516	17.857	0.000	0.983	0.017
U_1	0.062	0.561	14.876	0.000	0.998	0.002

ся вектором X_1, X_2, ..., X_m, $m > 1$. Задача состоит в том, чтобы по результатам измерений отнести объект к одной из нескольких групп (классов) G_1, G_2, ..., G_k, $k \geq 2$, т. е. нужно построить решающее правило, позволяющее по результатам измерений параметров объекта указать группу, к которой он принадлежит. Число групп заранее известно, также известно, что объект заведомо принадлежит к определенной группе.

Применительно к решаемой химической задаче цель анализа состояла в том, чтобы на основе известных свойств растворителей и катионов классифицировать константы устойчивости коронатов катионов щелочных металлов M[222]$^+$, иначе говоря, оптимальным способом отнести константы к одной из четырех групп (классов, кластеров), выявленных методами кластерного анализа.

В табл. 10 приведены итоги дискриминантного анализа. Достаточно малое значение общей Λ-Уилкса = 0.035; приближенное значение общего критерия Фишера $F_{набл}$(28, 151) = 32.119 и p < 0.000 свидетельствуют об успешности проведенной классификации методом дискриминантного анализа.

Λ-Статистика Уилкса (лямбда Уилкса) – это мера различий между 4 классами констант устойчивости по четырем дискриминантным переменным: ε, β, E_T, U_1. Существует несколько способов ее вычисления, один из них – расчет по формуле [19]:

где k – число уже вычисленных функций; символ

$$\Lambda = \prod_{i=k+1}^{g} \frac{1}{1+\lambda_i},$$

П означает, что для получения окончательного результата необходимо перемножить все члены; λ_i – собственные значения функции. Λ-статистика Уилкса учитывает, как различия между классами, так и когезивность или однородность каждого класса. Под когезивностью понимается степень скопления констант устойчивости (объектов) вокруг центроида их класса [19].

Величины Λ, близкие к нулю, говорят о высоком различении, т. е. центроиды классов хорошо разделены и сильно отличаются друг от друга по отношению к степени разброса констант устойчивости криптатов внутри классов. Увеличение Λ до ее максимального значения, равного 1, приводит к постепенному ухудшению различения, так как центроиды групп совпадают (нет групповых различий).

Значение частной Λ-Уилкса равно отношению лямбда Уилкса после добавления соответствующей переменной к лямбде Уилкса до добавления этой переменной. Частная лямбда характеризует единичный вклад соответствующей переменной в разделительную силу модели. Чем больше частная лямбда Уилкса, тем больше вклад переменной в общую дискриминацию [21, 22]. Из табл. 10 видно, что переменная ε дает наибольший вклад, переменная U_1 – вторая по значению вклада, переменная E_T – третья по значению вклада, а переменная β вносит наименьший вклад в общую дискриминацию.

F-исключить – это значения *F*-критерия, связанные с соответствующей частной лямбда Уилкса [21, 22]. Значения *p*-уровень – это уровни значимости критериев *F*-исключить. Значения p < 0.05 подтверждают статистическую значимость критериев *F*-исключить и желательность переменных в дискриминантной модели.

Таблица 11. Коэффициенты классифицирующих функций на основе дискриминантных переменных[a]

Переменная	G_1 *p* 0.453	G_2 *p* 0.078	G_3 *p* 0.188	G_4 *p* 0.281
ε, b_{k1}	0.092	0.346	0.143	0.140
β, b_{k2}	47.745	27.513	17.056	17.488
E_T, b_{k3}	39.641	70.209	26.004	26.986
U_1, b_{k4}	22.656	22.647	26.553	20.783
b_{k0}	–82.717	–109.613	–81.695	–54.592

[a] *p* – апостериорная (послеопытная) вероятность, пропорциональная количеству констант устойчивости криптатов в каждой группе (кластере).

Толерантность определяется как $(1 - R^2)$, где R – это коэффициент множественной корреляции данной переменной со всеми другими переменными в модели. Толерантность является мерой избыточности переменной в модели [21, 22]. Чем меньше значение толерантности, тем избыточнее переменная в модели, т.е. переменная несет малую дополнительную информацию. Формулы для толерантности, статистик *F*-включения и *F*-исключения довольно сложны [21], поэтому не приводятся.

На этой стадии дискриминантного анализа можно предположить, что электрические свойства растворителя (ε) и энергия ионизации атомов щелочных металлов (U_1) являются главными переменными, которые позволяют провести дискриминацию между различными классами криптатов $M[222]^+$.

Как следует из табл. 10, позиции четырех групп констант устойчивости сильно различаются по выбранным переменным, поэтому имеет смысл найти дискриминантные функции (классифицирующие функции) для каждой группы.

Р. Фишер [28] первым предположил, что классификация должна проводиться с помощью линейной комбинации дискриминантных переменных (предикторов), которая максимизирует различия между классами, но минимизирует дисперсию внутри классов. Разработка его предложения приводит к определению особой линейной комбинации для каждого класса, которая называется «классифицирующая функция» [21], и имеет вид (табл. 11):

$$f_k = b_{k0} + b_{k1}X_1 + b_{k2}X_2 + \ldots + b_{kp}X_p,$$

где f_k – значение функции для класса k, а b_{ki} – коэффициенты, которые необходимо определить, X_i – дискриминантные переменные: ε, β, E_T, U_1 (здесь $p = 4$ – число дискриминирующих переменных). Прогнозируемая по свойствам растворителя и катиона константа устойчивости криптата будет отнесена к классу с наибольшим значением *f*. Коэффициенты для классифицирующих функций определяются с помощью соотношения:

$$b_{ki} = (n_0 - g)\sum_{j=1}^{p} a_{ij}X_{jk},$$

где n_0 – общее количество констант устойчивости криптатов, иначе говоря наблюдений (64), в четырех группах g; b_{ki} – коэффициент для переменной i в выражении, соответствующему классу k; a_{ij} – элемент матрицы, обратной к внутригрупповой матрице сумм попарных произведений. Постоянный член рассчитывается по формуле:

$$b_{k0} = -0.5\sum_{j=1}^{p} b_{kj}X_{jk}.$$

Процедуры классификации могут использовать не только дискриминантные переменные, но и канонические дискриминантные функции, полученные с использованием алгоритмов канонического корреляционного анализа [21, 22].

Канонический анализ позволяет проанализировать природу различий между группами (кластерами). Согласно геометрической интерпретации анализа, дискриминантные переменные [21] – это оси *p*-мерного евклидова пространства. Каждый объект (наблюдение) является точкой этого пространства с координатами, представляющими собой наблюдаемые значения каждой переменной. Если классы отличаются друг от друга по наблюдаемым переменным, их можно представить скоплением точек в некоторых областях рассматриваемого пространства. Для определения положения класса вычисляется его центроид. Центроид класса является воображаемой точкой, координаты ко-

Таблица 12. Статистические показатели извлекаемых (ортогональных) дискриминантных функций (корней)

Извлеченокорней	Критерий Хи-квадрат последовательности извлечения корней					
	λ	*R*	Λ	χ	ν	*p*
0	3.952	0.893	0.035	197.805	12	0.000
1	2.245	0.832	0.173	103.420	6	0.000
2	0.779	0.662	0.562	33.971	2	0.000

[a] λ – собственное значение дискриминантной функции *D*, *R* – коэффициент канонической корреляции, Λ – значение статистики Λ Уилкса, χ^2 – значение статистики Хи-квадрат Пирсона, ν – число степеней свободы, *p* – уровень значимости соответствующего канонического корня.

торой есть средние значения переменных в данном классе [22].

В рассматриваемом случае константы устойчивости криптатов принадлежат 4-мерному пространству. Следовательно, четыре переменных – ε, β, E_T, U_1 определяют координаты центроида для каждого из четырех классов.

Каноническая дискриминантная функция является линейной комбинацией дискриминантных переменных. Она имеет следующее математическое представление [21, 22]:

$$D_{km} = u_0 + u_1 X_{1km} + u_2 X_{2km} + \ldots + u_p X_{pkm},$$

где D_{km} – значение канонической дискриминантной функции для *m*-го объекта в группе *k*; X_{ikm} – значение дискриминантной переменной X_i для *m*-го объекта в группе *k*; u_i – коэффициенты дискриминантных функций. Коэффициенты u_i для первой функции выбираются таким образом, чтобы ее средние значения для различных классов как можно больше отличались друг от друга [19, 21]. Коэффициенты второй функции выбираются так же, т. е. соответствующие средние значения должны максимально отличаться по классам, при этом налагается дополнительное условие, чтобы значения второй функции не коррелировали со значениями первой. Аналогично третья функция должна быть ортогональной первой и второй и т. д. Максимальное число дискриминантных функций, которое можно получить описанным способом, равно числу классов без единицы или числу дискриминантных переменных, в зависимости от того, какая из этих величин меньшая [19, 22].

В табл. 12 представлены результаты канонического анализа с пошаговым критерием χ^2 для канонических корней (Root) – канонических линейных дискриминантных функций *D* [19, 21–23].

Собственное (характеристическое) значение для каждой дискриминантной функции λ [19, 22] – это отношение межгрупповой суммы квадратов отклонений SS_B к внутригрупповой сумме квадратов отклонений SS_W. Большие собственные значения свидетельствуют о высокой статистической значимости извлеченных дискриминантных корней (функций).

Мощность вклада функции оценивается по критерию Хи-квадрат. Значение $p < 0.05$ указывает на статистически значимую мощность извлеченных дискриминантных функций [19]. Чем больше теоретические числа, рассчитанные на основе нулевой гипотезы (отсутствие различий между кластерами), будут отличаться от фактических, тем сильнее критерий Хи - квадрат будет отличаться от 0 (Λ Уилкса, наоборот, будет приближаться к 0), тем с большей вероятностью можно принять альтернативную статистическую гипотезу и говорить о статистической достоверности имеющихся различий в сравниваемых группах (кластерах) констант устойчивости криптатов. Величина χ^2 имеет Хи-квадрат распределение с $(p - k)(g - k - 1)$ степенями свободы [21].

$$\chi^2 = -\left\{n - \left[\frac{(p+g)}{2}\right] - 1\right\} \ln \Lambda_k,$$

где *k* – число извлеченных дискриминантных функций, равное $(g - 1)$.

Первая строка дает критерий значимости для всех дискриминантных функций (корней). Вторая строка содержит значимость дискриминантных функций, оставшихся после удаления первой функции и т. д. Таким образом, данные, приведенные в табл. 12 позволяет оценить, сколько значимых дискриминантных функций нужно интерпретировать. Как следует из табл. 12, статистически значимыми являются три дискриминантные функции.

Таблица 13. Коэффициенты канонических линейных классификационных функций D_{km} (Root)

Переменная	$D_{km} = u_0 + u_1 X_{1km} + u_2 X_{2km} + ... + u_p X_{pkm}$		
	Root 1	Root 2	Root 3
ε, u_1	–0.004	–0.045	0.001
β, u_2	7.205	3.281	0.266
E_T, u_3	4.411	–5.639	–0.085
U_1, u_4	–0.141	0.010	2.257
u_0	–5.221	2.746	–10.466
λ	3.952	2.245	0.779
Доля объясненнойдисперсии, %	56.700	88.800	100.000

Таблица 14. Координаты центроидов четырех групп констант устойчивости криптатов катионов

Группа	Средние канонических переменных		
	Root 1	Root 2	Root 3
G_1	1.843	0.780	0.021
G_2	1.592	–4.837	0.027
G_3	–2.143	0.059	1.502
G_4	–1.984	0.047	–1.044

В табл. 13 приведены коэффициенты u_i канонических линейных дискриминантных функций (корней) для исходных (нестандартизированных) свойств катионов и растворителей.

Первая дискриминантная функция Root 1 наиболее важная, так как отвечает за 56.7% объясненной дисперсии свойств растворителей и катионов. Вторая Root 2 отвечает за 32.1%, а третья Root 3 – за 11.2% объясненной дисперсии. Подставив в дискриминантные уравнения значения свойств растворителя и катиона (ε, β, E_T, U_1) можно рассчитать значения дискриминатных функций – D_{km}. Прогнозируемая константа устойчивости криптата, для которой рассчитаны Root 1, Root 2 и Root 3, будет отнесена к группе по минимальному расстоянию до соответствующего центроида группы (кластера). Координаты центроидов кластеров (средних значений канонических переменных) приведены в табл. 14.

По данным табл. 14 трудно судить о результатах разделения констант устойчивости криптатов по группам в многомерном пространстве переменных. Поэтому на рис. 3–5 приведены диаграммы рассеяния канонических значений констант устой-

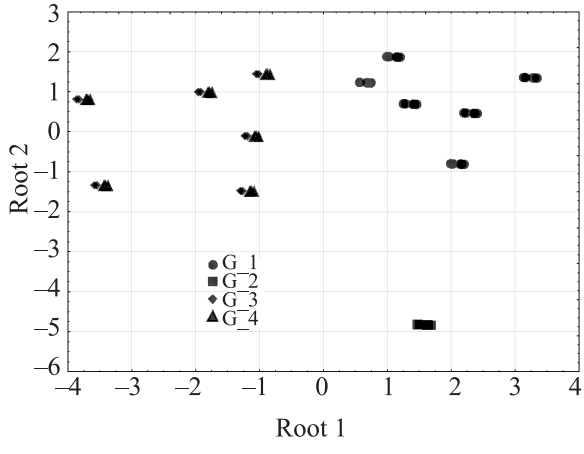

Рис. 3. Диаграмма рассеяния канонических значений констант устойчивости криптатов для пар значений дискриминантных функций Root 1–Root 2.

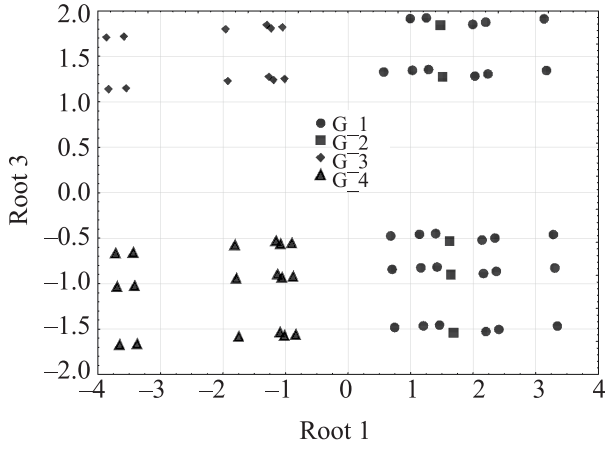

Рис. 4. Диаграмма рассеяния канонических значений констант устойчивости криптатов для пар значений дискриминантных функций Root 1–Root 3.

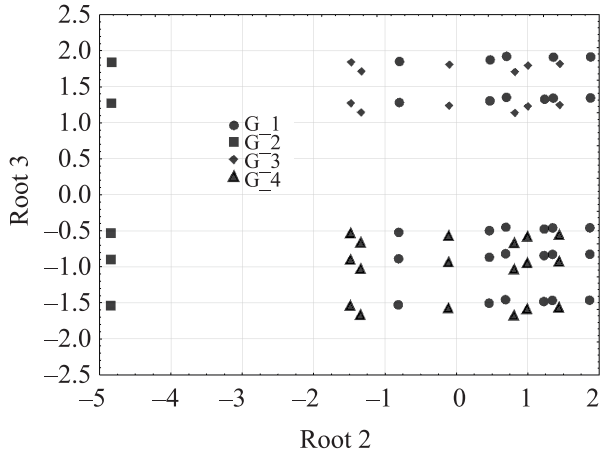

Рис. 5. Диаграмма рассеяния канонических значений констант устойчивости криптатов для пар значений дискриминантных функций Root 2–Root 3.

чивости криптатов для пар значений дискриминантных функций – Root.

Первая дискриминантная функция, определяющая 56.7% дискриминирующей мощности, в координатах Root 1 – Root 2 дискриминирует константы устойчивости криптатов между парами групп G_1, G_2 и G_3, G_4. Вторая дискриминантная функция (Root 2, 32.1% дискриминирующей мощности) дает основную дискриминацию между второй G_2 и первой G_1 группами. Третья дискриминантная функция в координатах Root 1 – Root 3 (Root 3, 11.2% дискриминирующей мощности) разделяет третью G_3 и G_4 группы. Вторая дискриминантная функция в координатах Root 2 – Root 3 идентифицирует константы устойчивости второй группы.

В табл. 15 приведены стандартизированные коэффициенты для канонических переменных. Стандартизованные коэффициенты применяются для выявления тех переменных (свойств), которые вносят наибольший вклад в значение дискриминантной функции. Чем больше абсолютная величина коэффициента, тем больше вклад этой переменной в значение дискриминантной функции.

Для первой дискриминантной функции вклад переменной β максимален, переменная E_T занимает второе место по значимости, остальные две переменные второстепенны. Для второй функции переменные ε, E_T вносят приблизительно одинаковый вклад, а переменная U_1 является доминантной для третьей функции.

Для выявления химической природы дискриминантных функций (взаимной зависимости отдельной переменной и дискриминантной функции) рассмотрим их корреляцию. Значения таких корреляций являются косинусами углов между векторами переменных и осями дискриминантных функций в многомерном пространстве [21]. Коэффициенты корреляции, называемые полными структурными коэффициентами, приведены в табл. 16.

Таблица 15. Стандартизованные дискриминантные коэффициенты

Переменная	Root 1	Root 2	Root 3
ε	–0.058	–0.607	0.017
β	0.903	0.411	0.033
E_T	0.506	–0.646	–0.010
U_1	–0.062	0.004	0.999
λ	3.952	2.245	0.779
Доля объясненнойдисперсии	0.567	0.888	1.000

Таблица 16. Полные структурные коэффициенты

Переменная	Корреляции: переменные – канонические дискриминантные функции (обобщенные внутригрупповые корреляции)		
	Root 1	Root 2	Root 3
ε	0.113	–0.618	0.010
β	0.864	0.369	0.052
E_T	0.444	–0.732	0.031
U_1	–0.026	–0.009	0.999

Таблица 17. Классификационная матрица дискриминантного анализа[a]

Группы	Точностьпредсказания, %	G_1 *p* 0.453	G_2 *p* 0.078	G_3 *p* 0.188	G_4 *p* 0.281
G_1	100.0	29	0	0	0
G_2	100.0	0	5	0	0
G_3	100.0	0	0	12	0
G_4	100.0	0	0	0	18
Всего	100.0	29	5	12	18

[a] Строки – наблюдаемая классификация, колонки – предсказанная классификация.

Если абсолютная величина такого коэффициента велика, вся информация о дискриминантной функции заключена в этой переменной. Если же коэффициент близок к нулю – их зависимость мала. Таким образом, коэффициенты корреляции, приведенные в табл. 16, свидетельствуют о том, что дискриминирующая мощь первой дискриминантной функции преимущественно определяется электронодонорными свойствами растворителей β (0.864), а третьей – свойствами катионов U_1 (0.999). Разделяющая мощь второй дискриминантной функции обусловлена главным образом электрическими ε (–0.618) и электроакцепторными E_T (–0.732) свойствами растворителей.

Таблица 17 представляет собой классификационную матрицу, которая позволяет говорить о точности дискриминантной процедуры, количестве правильно классифицированных констант устойчивости криптатов и тем самым косвенно подтвердить степень разделения классов. Четыре переменных правильно предсказывают распределение по группам всех констант устойчивости криптатов. Точность предсказания в этом случае – 100% (сумма правильных предсказаний 64, поделенная на общее число наблюдаемых констант устойчивости – 64). Процент наблюдаемых констант устойчивости, которые были классифицированы правильно, является дополнительной мерой различий между группами [21, 22].

Таким образом, дискриминантная модель классификации констант устойчивости криптатов M[222]$^+$ по свойствам растворителей и катионов на 100% подтвердила результаты кластеризации констант устойчивости методом *k*-средних.

Деревья классификации представляют собой последовательные иерархические структуры, состоящие из узлов, которые содержат правила, т. е. логические конструкции вида «если …, то …». Конечными узлами дерева являются «листья», соответствующие найденным решениям и объединяющие некоторое количество объектов (наблюдений) в группы (классы) [19, 29].

Химическая задача состояла в построении дерева классификации констант устойчивости криптатов катионов щелочных металлов (зависимая категориальная переменная, характеризуемая четырьмя группами) по четырем свойствам растворителей ε, β, E_T и катионов U_1 (независимые переменные в порядковой шкале).

Процесс построения дерева классификации состоял из четырех основных этапов [19]: (1) выбор критерия точности прогноза, (2) выбор вариантов ветвления, (3) определение момента, когда дальнейшие ветвления следует прекратить, (4) определение «подходящего размера» дерева.

Цель анализа с помощью деревьев классификации заключалась в том, чтобы получить максимально точный прогноз (первый этап). Наиболее точным прогнозом считается такой, который связан с наименьшей ценой ошибки классификации. В программе STATISTICA [19] под ценой ошибки классификации понимается доля неправильно классифицированных наблюдений – неправильных распределений констант устойчивости криптатов в группы, которые, как отмечено ранее, были сформированы методом *k*-средних кластерного анализа. Как правило, самый лучший прогноз – такой, который дает наименьший процент неправильных классификаций.

В работе выбран вариант анализа, когда цена ошибки классификации для всех классов одинаковая (Equal); все внедиагональные элементы

Таблица 18. Структура дерева классификации

Вершина	Дочерние вершины, наблюдаемые, предсказанные классы, условия ветвления								
	левая ветвь	правая ветвь	Класс 1	Класс 2	Класс 3	Класс 4	предсказанный класс	значение переменной ветвления	ветвление по переменной
1	2	3	29	5	12	18	1	0.57	β
2	4	5	0	5	12	18	4	4.739	U_1
3			29	0	0	0	1		
4	6	7	0	3	0	18	4	0.74	E_T
5	8	9	0	2	12	0	3	0.74	E_T
6			0	0	0	18	4		
7			0	3	0	0	2		
8			0	0	12	0	3		
9			0	2	0	0	2		

матрицы цен ошибок классификации (прогнозируемые классы – по строкам, наблюдаемые классы – по столбцам) принимались равными 1; в выбранные значения априорных вероятностей (Prior probabilities) для всех классов зависимой переменной поправки не вводились.

Второй этап анализа заключался в том, чтобы выбрать способ ветвления по значениям предикторных переменных (свойств растворителей и катионов), Ветвления последовательно начинаются с корневой вершины, затем переходят к вершинам потомкам, пока дальнейшее ветвление не прекратится и «неразветвленные» вершины потомки станут терминальными. Терминальные вершины (или листья) – это узлы дерева, начиная с которых никакие решения больше не принимаются. Началом дерева считается самая верхняя решающая вершина, которую иногда также называют корнем дерева [19, 23].

Выбран тип ветвления C&RT (Classification and Regression Trees) – полный перебор вариантов одномерного ветвления методом C&RT (Style Exhaustive Search for Univariate Splits). Этот метод можно использовать для всех типов предикторных переменных. В отличие от дискриминантных методов ветвления, в методе C&RT, для того чтобы найти наилучший вариант ветвления, проводится последовательный перебор всех возможных комбинаций уровней предикторных переменных. Количество уровней, образующихся от узлов, не считая корневую вершину, характеризуют глубину дерева [29].

В качестве критерия согласия была выбрана мера Джини (Gini measure) [19]. Критерии согласия используются для выбора наилучшего из всех возможных вариантов ветвления. Мера Джини однородности вершины принимает нулевое значение, когда в данной вершине имеется всего один класс.

Третий этап анализа заключался в выборе момента, когда следует прекратить дальнейшие ветвления. Выбран вариант остановки: отсечение по ошибке классификации (Prune on Misclassification Error) [19, 29].

С определением момента, когда дальнейшие ветвления следует прекратить, непосредственно связан четвертый этап – определение «подходящих размеров» дерева. Очевидно, что чем больше размерность дерева классификации, тем точнее прогноз.

В табл. 18 представлены номера вершин (node); номера дочерних вершин (child nodes) на левой и правой ветвях (left, right branch); исходное количество объектов (observed) в классах; предсказанные классы (predicted classes); условия ветвления (split conditions).

Из табл. 18 следует, что левая ветвь содержит четыре узла под номерами 2, 4, 6, 8; правая – четыре узла под номерами 3, 5, 7, 9. Пять вершины 3, 6, 7, 8 и 9 являются терминальными. Из строки 1 таблицы вытекает, что в первой вершине все константы устойчивости криптатов классифицированы (предсказаны) как Класс 1 (обозначения классов в правом верхнем углу вершин рис. 6), по

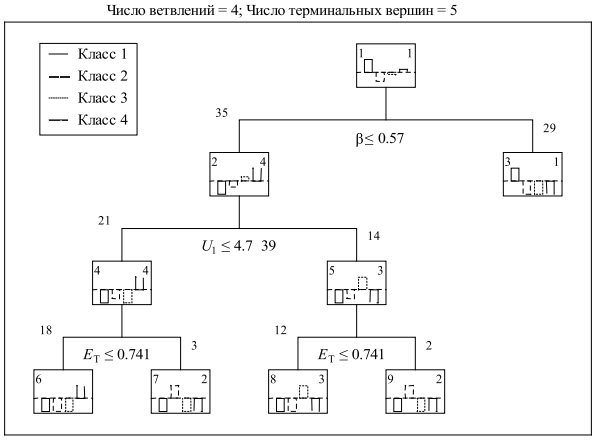

Рис. 6. Граф дерева классификации констант устойчивости криптатов $M[222]^+$.

наибольшему числу констант в первом классе (29). Из вершины 1 выходят две ветви (правая и левая) с соответствующими вершинами 2 и 3 (обозначения в левом верхнем углу вершин).

Условие (правило) разделения констант устойчивости по вершинам 2 и 3 следующее: если значение переменной $\beta \le 0.57$, то константы устойчивости классифицируются как Класс 4, в противном случае – как Класс 1. Из строк 2 и 3 следует, что по данному правилу 5, 12, и 18 (всего 35) наблюдаемых констант устойчивости Классов 2, 3 и 4 предсказаны как Класс 4, а 29 констант устойчивости Класса 1 правильно классифицированы как Класс 1. Из вершины 2 также выходят две ветви (правая и левая) с соответствующими вершинами 4 и 5. Условие разделения констант устойчивости по вершинам 4 и 5 подчиняется правилу: если $U_1 \le 4.739$, то константы устойчивости классифицируются как Класс 4, в противном случае – как Класс 3. Из строк 4 и 5 таблицы вытекает, что по данному правилу 3 и 18 (всего 21) наблюдаемых констант устойчивости Классов 2 и 4 предсказаны как Класс 4, а 2 и 12 (всего 14) наблюдаемых констант устойчивости Классов 2 и 3 предсказаны как Класс 3.

Дальнейшая интерпретация результатов табл. 18 значительно упрощается, если воспользоваться графом дерева классификации, приведенным на рис. 6.

В табл. 19 приведены результаты деревьев классификации для правила остановки ветвления – отсечение по ошибке классификации констант устойчивости криптатов (выбранное дерево классификации отмечено *звездочкой*) [19, 29].

В табл. 20 приведена матрица ошибок классификации глобальной кросс-проверки [19, 29]. Из данной таблицы следует, что при глобальной кросс-проверке две константы устойчивости Класса 3 неверно классифицированы как Класс 2, все остальные константы устойчивости классифицированы верно. При этом цена глобальной кросс-проверки (Global CV cost) составила 0.031250, стандартное отклонение (s.d. Global CV cost) цены – 0.02175 и эти величины совпадают с

Таблица 19. Статистика для последовательности деревьев классификации

Номер вершины	Терминальные вершины	Цена кросс-проверки	Стандартная ошибка	Цена обучения	Сложность усеченного дерева
1*	5	0.031250	0.021749	0.000000	0.000000
2	4	0.062500	0.030258	0.031250	0.031250
3	3	0.078125	0.033546	0.078125	0.046875
4	2	0.265625	0.055208	0.265625	0.187500
5	1	0.546875	0.062225	0.546875	0.281250

Таблица 20. Матрица ошибок классификации глобальной кросс-проверки[a]

Класс	Класс 1	Класс 2	Класс 3	Класс 4
1		0	0	0
2	0		0	0
3	0	2		0
4	0	0	0	

[a] Матрица: предсказанные ошибки (строки) × наблюдаемые ошибки (колонки); цена глобальной кросс-проверки = 0.03125; стандартное отклонение цены = 0.02175.

Таблица 21. Статистические показатели прямой и обратной пошаговой регрессии[a]

Число наблюдений	Beta	Стандартная ошибка Beta	*b*	Стандартная ошибка *b*	t(61)	*p*-Уровень
b_0			12.85	1.06	12.15	0.000
β	–0.46	0.10	–6.06	1.39	–4.37	0.000
ε	–0.35	0.10	–0.07	0.02	–3.39	0.001

[a] R 0.578, R^2 0.334, R^2_a 0.313, F(2.61) 15.32 $p < 0.000$, стандартная ошибка аппроксимации – 2.78.

Таблица 22. Показатели частной корреляции

Переменная	Beta	Частная корреляция	Получастная корреляция	Толерантность	R^2	$t_{набл}(61)$	*p*-Уровень
β	–0.46	–0.488	–0.456	1.0	0.000	4.37	0.000
ε	–0.35	–0.399	–0.354	1.0	0.000	3.39	0.001

ценой кросс-проверки (табл. 19). Таким образом, процедура классификации констант устойчивости криптатов методом Дерева классификации проведена успешно и ее результаты на 96.9% (62/64) подтвердили результаты кластерного, дискриминантного и канонического анализа.

Получено решающее правило, состоящее из четырех этапов (табл. 18, рис. 6), которое произвольную (прогнозируемую) константу устойчивости криптата относит к одному из четырех классов по значениям свойств растворителя (β, E_T) и катиона (U_1).

Множественная линейная регрессия. Краткое описание модуля множественная регрессия (Multiple Regression) [19, 29] в программе STATISTICA 12, применительно к рассматриваемой задаче аппроксимации констант устойчивости криптатов катионов по свойствам растворителей и катионов: Y_i – наблюдаемые значения константы устойчивости криптатов lgK_i; PrY_i – предсказанные значения (predictable values) lg$K$, вычисленные по уравнению регрессии: $PrY_i = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n$, где X_i – свойства растворителей и катионов, b_i – коэффициенты уравнения регрессии, $i = 1, 2, \ldots, n$), n – количество переменных (ε, β, E_T, U_1), b_0 – свободный член; $Res = (Y_i - PrY_i)$ – остатки (residuals), разность между наблюдаемыми значениями lg$K$ и предсказанными; $SS = \Sigma_i(Y_i - \bar{Y})^2$ – сумма квадратов Y_i, скорректированная на среднее, где среднее $\bar{Y} = \Sigma_i Y_i/n$; $SSPr = \Sigma_i(PrY_i - \bar{Y})^2$ – сумма квадратов PrY_i, скорректированная на среднее; $SSRes = \Sigma_i(PrY_i - \bar{Y})^2$ – сумма квадратов остатков; $R^2 = (1 - SSRes/SS)$ – коэффициент детерминации; $R = \sqrt{R^2}$ – коэффициент множественной корреляции, характеризует тесноту связи между предикторами и константой устойчивости криптата (откликом), а также является оценкой качества предсказания, изменяется в пределах от 0 до 1; $R^2_a = 1 - (1 - R^2)[n/(n - k)]$ – скорректированное (adjusted) значение R^2, где k – число параметров (коэффициентов b_i) в регрессионном уравнении без учета свободного члена.

Для получения регрессионных уравнений (математических моделей) использован метод пошаговой регрессии. Эта процедура вводит или выводит предикторы (свойства растворителей и катионов) из уравнения регрессии по очереди, основываясь на серии F-тестов, t-тестов или других подходах [19].

В табл. 21 представлены результаты прямой и обратной пошаговой регрессии. Таблица содержит стандартизованные (*Beta*) и нестандартизованные (b) регрессионные коэффициенты, их стандартные ошибки и уровни значимости. Коэффициенты *Beta* оцениваются по стандартизованным данным, имеющим выборочное среднее, равное 0, и стандартное отклонение, равное 1. Близкие значения коэффициентов *Beta* позволяют заключить, что вклады каждого предиктора (β и ε) в предсказание константы устойчивости криптатов практически одинаковые. Отрицательный знак коэффициентов при этих переменных означает, что с увеличением значений β и ε, устойчивость комплексов катионов с криптандом [222] уменьшается. Коэффициенты уравнения регрессии b_1, b_2 и свободный член b_0 статистически значимы при уровне значимости p 0.05, так как $p < 0.05$.

Таблица 22 содержит коэффициенты *Beta*, частные коэффициенты корреляции, получастные коэффициенты корреляции (Semipart Cor), толерантности (Tolerance), коэффициенты детерминации (*R*-square), значения t-критерия и уровни значимости *p* – вероятности отклонения гипотезы о значимости частного коэффициента корреляции.

Частные коэффициенты корреляции (Partial Cor) показывают степень влияния одного предиктора на константу устойчивости криптата (отклик) в предположении, что остальные предикторы закреплены на постоянном уровне, т. е. контролируется их влияние на отклик [19, 29]. Из табл. 22 следует, что возрастание электронодонорной способности растворителей в большей степени влияет на снижение устойчивости криптатов $M[222]^+$, чем возрастание диэлектрической проницаемости.

Получастная корреляция – корреляция предиктора и константы устойчивости криптата в предположении, что контролируется влияние других предикторов на данный предиктор, но не контролируется влияние предикторов на отклик [19, 29] (константу устойчивости комплексов $M[222]^+$). Если получастная корреляция мала, в то время как частная корреляция относительно велика, то соответствующий предиктор может иметь самостоятельную часть в объяснении изменчивости константы устойчивости (зависимой переменной), т. е. часть, которая не объясняется другими предикторами. Из данных табл. 22 видно, что предикторы β и ε не имеют самостоятельной части в объяснении изменчивости устойчивости криптатов, так как их частные и получастные корреляции достаточно близки.

Коэффициент детерминации – квадрат коэффициента множественной корреляции между данной переменной и всеми остальными переменными, входящими в уравнение регрессии. Из таблицы следует, что коэффициенты детерминации близки к нулю, что свидетельствует об ортогональности предикторов β и ε по отношению к другим свойствам растворителей и катионам. Толерантность $(1 - R^2)$, как мера избыточности переменной, подтверждает этот вывод. $t_{набл}(61)$ – значение критерия Стьюдента для проверки гипотезы о значимости частного коэффициента корреляции с указанным (в скобках) числом степеней свободы. Из табл. 22 следует, что $t_{набл}(61) > t_{кр}(61) = 2.00$ – это означает, что коэффициенты корреляции статистически значимы для переменных β и ε.

Значение коэффициента множественной корреляции $R = 0.578$ (табл. 21) свидетельствует о том, что построенная регрессионной модель обладает недостаточной прогностической мощностью для предсказания устойчивости криптатов $M[222]^+$ по свойствам растворителей и катионов. Поэтому были привлечены нейросетевые технологии.

Нейросетевой анализ. Процесс обучения нейронной сети [19, 30] заключался в подстройке ее внутренних параметров под конкретную задачу [19] – построение нейросетевого аппроксиматора и классификаторов констант устойчивости криптатов по свойствам растворителей и катионов. Алгоритм работы нейронной сети является итеративным, его шаги называют эпохами или циклами. Эпоха – одна итерация в процессе обучения, включающая предъявление к обучению всех наблюдений (примеров) из обучающего множества. Сеть обучалась на выборке (train), включающей 70% наблюдений, процесс обучения контролировался (test) на контрольной выборке (15% наблюдений), обученная сеть проверялась на проверочной (validation) выборке (15% наблюдений). Контрольная выборка используется для остановки обучения в момент наилучшей обучающей способности нейронной сети (минимальная ошибка на контрольной выборке). Проверочная выборка не участвует в обучении вообще, после завершения обучения она используется для оценки производительности полученной сети.

В табл. 23 приведены основные характеристики лучшего (из 1000 обученных) нейросетевого аппроксиматора – многослойного персептрона MLP 4-7-1. Архитектура MLP 4-7-1 обозначает: многослойный персептрон с 4-мя входными и 1-й выходной переменными, и тремя слоями: входной – 4 нейрона, промежуточный – 7 нейронов и выходной – 1 нейрон.

Коэффициенты корреляции на обучающем, контрольном и тестовом множествах равны 0.9897, 0.9851 и 0.9930 соответственно. Статистические характеристики обученной нейросетевой модели персептронного типа MLP 4-7-1 (табл. 23) отражают успешность проведенного обучения. Так, качество обучения на различных множествах больше

Таблица 23. Итоги обучения нейросетевого аппроксиматора MLP 4-7-1[a]

Архитектура	Производительность обучения	Контрольная производительность	Тестовая производительность	Ошибка обучения	Контрольная ошибка	Тестовая ошибка	Алгоритм обучения	Функция ошибки	Функция активации скрытых нейронов	Функция активации выходных нейронов
MLP 4-7-1	0.990	0.993	0.985	0.105	0.160	0.117	BFGS 112	SOS	Lgistic	Identity

[a] Производительность обучения, контрольная производительность, тестовая производительность – отношение стандартного отклонения ошибки прогноза к стандартному отклонению исходных данных на соответствующих выборках; Ошибка обучения, контрольная ошибка, тестовая ошибка – ошибки сети на соответствующих выборках; BFGS – алгоритм Бройдена–Флетчера–Гольдфарба–Шанно [31, 32]; SOS – среднеквадратичная ошибка $E = \frac{1}{P}\sum_{i=1}^{P}(\lg K_{\text{расч},i} - \lg K_{\text{эксп},i})^2$, P – количество обработанных примеров в выборке; Identity – тождественная $\varphi(x) = x$, Lgistic – логистическая $\varphi(x) = 1/(1 + \exp(-tx))$ [19].

Таблица 24. Итоги кластеризации констант устойчивости криптатов $M[222]^+$ многослойным персептроном MLP 4-7-4

Архитектура	Показатели кластеризации	Класс 1	Класс 2	Класс 3	Класс 4	Все
MLP 4-7-4	Все	29	5	12	18	64
	Правильно	29	5	12	18	64

98%, ошибка обучения на обучающем множестве 0.105, на контрольном 0.160, на тестовом – 0.117. Эти данные также свидетельствуют о том, что нейросетевая модель обладает большей прогнозирующей силой, чем модель множественной линейной регрессии, коэффициент корреляции которой 0.578.

Обученный нейросетевой классификатор MLP 4-7-4 (табл. 24) имеет следующие основные характеристики: производительность обучения – 100%, контрольная производительность – 100%, тестовая производительность – 100%; алгоритм обучения – BFGS 8; функция ошибки Entropy - кросс-энтропийные потери

$$H(p,q) = -\sum_x p(x)\lg q(x),$$

p и q – несвязанные друг с другом случайные переменные [19]; функции активации нейронов: скрытых – Identity $\varphi(x) = e^x$; выходных – Softmax

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}},$$

где $\vec{z}$ – входной вектор функции softmax, z_i – элементы входного вектора, K – количество классов в мультиклассификаторе [19, 33].

Алгоритм многослойного персептрона MLP 4-7-4 на 100% подтвердил правомочность кластеризации методом k-средних (табл. 24).

По степени влияния на классификацию (группирование) констант устойчивости криптатов свойства растворителей и катионов располагаются в следующий ряд: $\beta(1.33) > U_1(1.13) > \varepsilon(0.90) > E_T(0.79)$. Техника проведения анализа чувствительности состоит в изменении выбранных параметров в определенных пределах, при условии, что остальные параметры остаются неизменными [19]. Таким образом, электронодонорные свойства растворителей и энергия ионизации атомов щелочных металлов – наиболее критические пере-

Таблица 25. Итоги кластеризации констант устойчивости криптатов $M[222]^+$ самоорганизующейся сетью Кохонена SOFM 8-4

Сеть	Ошибка			Алгоритм обучения	Количественный состав кластеров			
	обучающая выборка, 70%	контрольная выборка, 15%	тестовая выборка, 15%		Класс 1	Класс 2	Класс 3	Класс 4
	$M[222]^+$, M = Li, Na, K, Rb или Cs							
SOFM 8-4	0.1578	0.2547	0.2008	Kohonen 1000	29	5	12	18

Таблица 26. Наблюдаемые ($\lg K_{эксп}$) и аппроксимированные ($\lg K_{MLP}$) персептроном MLP 4-7-1 значения констант комплекообразования ($\lg K$) катионов с криптандом [222]

Растворитель	$\lg K_{эксп}$ $Li[222]^+$	$\lg K_{MLP}$ $Li[222]^+$	$\lg K_{эксп}$ - $Na[222]^+$	$\lg K_{MLP}$ $Na[222]^+$	$\lg K_{эксп}$ $K[222]^+$	$\lg K_{MLP}$ $K[222]^+$	$\lg K_{эксп}$ $Rb[222]^+$	$\lg K_{MLP}$ $Rb[222]^+$	$\lg K_{эксп}$ $Cs[222]^+$	$\lg K_{MLP}$ $Cs[222]^+$
Вода	0.98	0.62	3.98	4.92	5.47	5.82	4.24	4.29	1.47	1.15
Метанол	2.59	3.22	7.98	7.97	10.41	9.98	9.10	8.15	4.00	4.15
Этанол	2.57	3.18	8.57	7.99	10.50	10.48	9.28	8.72	4.17	4.82
н-Пропанол	2.49	3.00	8.39	7.85	10.80	10.58	9.09	9.06	4.55	5.73
Ацетон	4.62	5.11	8.89	8.43	10.04	10.31	8.39	8.37	3.96	4.16
Пропиленкарбонат	6.94	7.22	10.54	10.73	11.19	11.00	9.02	8.85	4.00	4.37
N-Метилпирролидон	2.97	2.31	5.83	6.56	8.41	8.82	7.28	7.02	4.38	3.08
ДМСО	1.05	0.75	5.32	5.02	7.11	7.17	5.85	5.60	2.19	2.23
ДМФА		2.80*	6.17	6.88	7.98	8.74	6.78	6.84	2.16	2.71
Нитрометан	11.47	11.06	13.56	13.34	12.58	12.70	10.30	10.28	5.10	5.43
Ацетонитрил	6.98	6.26	9.63	9.78	11.01	11.29	9.50	9.26	4.57	4.87
Бензонитрил	9.14	8.51	11.56	11.78	13.06	13.11	11.00	10.92	6.59	6.24
1,2-Дихлорэтан	7.90	8.58	10.60	10.59	13.00	13.08	12.49	11.54	8.50	8.26

менные, которые в наибольшей степени влияют на осуществимость и эффективность разделения констант устойчивости криптатов катионов на четыре класса.

В табл. 25 приведены основные характеристики самоорганизующего классификатора SOFM 8-4, на 100 % подтвердившего результаты кластеризации методом *k*-средних.

Аппроксимирующие и прогностические возможности персептрона MLP 4-7-1. В табл. 26 приведены результаты применения обученного персептрона MLP 4-7-1 для аппроксимации зависимости экспериментальных констант устойчивости криптатов катионов щелочных металлов от свойств растворителей и катионов.

В табл. 27 приведены результаты прогнозирования обученным персептроном MLP 4-7-1 констант устойчивости криптата $K[222]^+$ по свойствам смешанных растворителей вода-ацетонитрил и катиона калия. Экспериментальные константы комплексообразования $\lg K_{эксп}$, взятые из работы [34], не использовались в обучении нейронной сети.

В табл. 28 приведены прогнозные $\lg K_{MLP}$ и литературные $\lg K_{эксп}$ значения констант устойчивости криптатов $Na[222]^+$ и $K[222]^+$ в органических растворителях [35].

Из анализа прогнозных данных, приведенных в табл. 27 и 28, вытекают возможные пути повышения прогностической мощности нейросетевой модели MLP 4-7-1: (1) введение в обучающую выборку данных по константам комплексообразования $M[222]^+$ в смешанных растворителях разного состава (табл. 27) и (2) пополнение обучающей выборки паттерновыми растворителями, свойства которых изменяются в широких пределах.

Обучение персептрона-аппроксиматора MLP 4-7-1 проведено на свойствах растворителей, диэлектрическая проницаемость которых изменялась

Таблица 27. Наблюдаемые ($\lg K_{эксп}$) [34] и предсказанные ($\lg K_{MLP}$) персептроном MLP 4-7-1 значения констант устойчивости криптатов $K[222]^+$ в смешанных растворителях вода–ацетонитрил

Мол. доля MeCN	ε	β	E_T	U_1	$\lg K_{эксп}$ $K[222]^+$	$\lg K_{MLP}$ $K[222]^+$	Остатки
H_2O	78.36	0.47	1.000	4.34	5.60	5.82	–0.22
0.1	70.47	0.34	0.890	4.34	6.50	8.32	–1.82
0.2	62.29	0.39	0.830	4.34	7.10	8.48	–1.38
0.3	55.69	0.41	0.810	4.34	7.70	8.80	–1.10
0.4	50.78	0.40	0.790	4.34	8.10	9.19	–1.09
0.5	47.06	0.40	0.770	4.34	8.60	9.33	–0.73
0.6	44.02	0.42	0.750	4.34	8.90	9.29	–0.39
0.7	41.42	0.44	0.730	4.34	9.20	9.26	–0.06
0.8	39.26	0.43	0.700	4.34	9.70	9.34	0.36
0.9	37.60	0.40	0.640	4.34	10.30	9.67	0.63
MeCN	35.86	0.37	0.460	4.34	11.40	11.80	–0.40

Таблица 28. Наблюдаемые ($\lg K_{эксп}$) [35] и предсказанніе ($\lg K_{MLP}$) персептроном MLP 4-7-1 значения констант устойчивости криптатов $Na[222]^+$ и $K[222]^+$ в органических растворителях

Криптат	Растворитель	ε	β	E_T	U_1	$\lg K_{эксп}$ $M[222]^+$	$\lg K_{MLP}$ $M[222]^+$	Остатки
$Na[222]^+$	Тетраметиленсульфон	43.26	0.39	0.410	5.14	10.50	11.79	–1.29
$K[222]^+$	Тетраметиленсульфон	43.26	0.39	0.410	4.34	11.30	12.79	–1.49
$Na[222]^+$	Формамид	109.50	0.48	0.775	5.14	6.20	–0.21	6.41
$K[222]^+$	Формамид	109.50	0.48	0.775	4.34	7.90	3.38	4.52
$Na[222]^+$	N,N-Диметилацетамид	37.78	0.76	0.377	5.14	5.70	6.12	–0.42
$K[222]^+$	N,N-Диметилацетамид	37.78	0.76	0.377	4.34	8.00	8.23	–0.23

в пределах от 10.36 ($C_2H_4Cl_2$) до 78.36 (H_2O). Диэлектрическая проницаемость формамида – 109.50 (табл. 28). Это одна из причин неудовлетворительного прогнозирования констант устойчивости криптатов в формамиде.

Представлена методология кластеризации (группирования) и прогнозирования устойчивости криптатов катионов щелочных металлов по свойствам растворителей и катионов на основе компьютерного анализа экспериментальных данных по константам устойчивости комплексов состава 1:1 криптанда [2.2.2] с катионами Li^+, Na^+, K^+, Rb^+, Cs^+.

Методологическую основу классификации и прогнозирования устойчивости криптатов составляет сочетание алгоритмов разведочных методов анализа, множественной линейной регрессии, многослойных искусственных нейронных сетей и самоорганизующихся сетей Кохонена. Совместное использование разведочных методов анализа дает возможность обосновать статистическую значимость свойств растворителей и катионов, влияющих на устойчивость криптатов катионов щелочных металлов в разных растворителях и осуществить отбор дескрипторов для построения нейросетевых прогнозных моделей аппроксимации и классификации устойчивости катионных комплексов $M[2.2.2]^+$.

Построенные разведочные и нейросетевые модели позволяют предсказывать константы устойчивости коронатов катионов щелочных металлов в органических и водно-органических средах по свойствам растворителей и катионов, а также оптимизировать планирование экспериментов в растворителях, в которых комплексообразование криптандов с катионами еще не изучено, либо исследовано недостаточно полно.

КОНФЛИКТ ИНТЕРЕСОВ

Автор заявляет об отсутствии конфликта интересов.

СПИСОК ЛИТЕРАТУРЫ

1. *Pedersen C.J.* // Science. 1988 Vol. 241. N 4865. P. 536. doi 10.1126/science.241.4865.536
2. *Lehn J.-M.* // Angew. Chem. Int. Ed. Engl. 1988. Vol. 27. N 1. P. 89. doi 10.1002/anie.198800891
3. *Cram D.J.* // J. Incl. Phenom. Macrocycl. Chem. 1988. Vol. 6. N 4. P. 397. doi 10.1007/bf00658982
4. *Лен Ж.-М.* Супрамолекулярная химия: Концепции и перспективы. Новосибирск: Наука. Сиб. предприятие РАН, 1998. 334 с.
5. *Стид Дж.В., Этвуд Дж.Л.* Супрамолекулярная химия. М.: ИКЦ «Академкнига», 2007. Т. 1. 2007. 480 с.
6. *Цивадзе А.Ю., Ионова Г.В., Михалко В.К., Кострубов Ю.Н.* // Усп. хим. 2007. Т. 76. № 3. С. 237; *Tsivadze A.Yu, Ionova G.V., Mikhalko V.K., Kostrubov Yu.N.* // Russ. Chem. Rev. 2007. Vol. 76. N 3. P. 213. doi 10.1070/RC2007v076n03ABEH003628
7. *Соловьев И.П.* Дис. ... докт. хим. наук. М., 2007. 350 с.
8. *Бондарев Н.В.* Термодинамика равновесий. Эффекты среды и нейросетевой анализ. Saarbrucken: LAP LAMBERT Academic Publishing, 2012. 380 с.
9. *Бондарев Н.В.* // ЖОХ. 2019. Т. 89. № 2. С. 288. doi 10.1134/S0044460X19020197; *Bondarev N.V.* // Russ. J. Gen. Chem. 2019. Vol. 89. N 2. P. 281. doi 10.1134/S1070363219020191
10. *Бондарев Н.В.* // ЖОХ. 2019. Т. 89. № 7. С. 1085. doi 10.1134/S0044460X1907014X; *Bondarev N.V.* // Russ. J. Gen. Chem. 2019. Vol. 89. N 7. P. 1438. doi 10.1134/S1070363219070144
11. *Бондарев Н.В.* // ЖОХ. 2020. Т. 90. № 6. С. 953. doi 10.31857/S0044460X20060170; *Bondarev N.V.* // Russ. J. Gen. Chem. 2020. Vol. 90. N 6. P. 1040. doi 10.1134/S1070363220060171
12. *Бондарев Н.В.* // ЖОХ. 2020. Т. 90. № 8. С. 1272; *Bondarev N.V.* // Russ. J. Gen. Chem. 2020. Vol. 90. N 8. P. 1476. doi 10.1134/S1070363220080149
13. *Бондарев Н.В.* // ЖОХ. 2020. Т. 90. № 10. С. 1583. doi 10.31857/S0044460X20100145; *Bondarev N.V.* // Russ. J. Gen. Chem. 2020. Vol. 90. N 10. P. 1906. doi 10.1134/S107036322010014X
14. *Marcus Y.* // Rev. Anal. Chem. 2004. Vol. 23. N 4. P. 269. doi 10.1515/REVAC.2004.23.4.269
15. *Marcus Y.* The Properties of Solvents. Chichester: John Wiley & Sons. 1999. Vol. 4. 399 p.
16. *Shannon R.D., Prewitt C.T.* // Acta Crystallgr. (B). 1969. Vol. 25. N 5. P. 925. doi 10.1107/s0567740869003220
17. Таблицы физических величин. Справочник / Под ред. И.К. Кикоина. М.: Атомиздат, 1976. 1008 с.
18. Физические величины. Справочник / Под ред. И.С. Григорьева, Е.З. Мейлихова. М.: Энергоатомиздат, 1991. 1232 с.
19. StatSoft – Электронный учебник по статистике. http://statsoft.ru/home/textbook/
20. *Лемешко Б.Ю.* Критерии проверки отклонения распределения от нормального закона. Руководство по применению. 2014. Новосибирск: НГТУ. 192 с.
21. *Ким Дж.-О., Мьюллер Ч.У., Клекка У.Р.* Факторный, дискриминантный и кластерный анализ. М.: Финансы и статистика. 1989. 216 с.
22. *Малхорта Н.К.* Маркетинговые исследования. Практическое руководство. М.: Издательский дом «Вильямс», 2002. 960 с.
23. *Боровиков В.П.* STATISTICA. Искусство анализа данных на компьютере: Для профессионалов. СПб: Питер, 2003. 686 с.
24. *Наследов А.* IBM SPSS Statistics 20 и AMOS: профессиональный статистический анализ данных. СПб: Питер, 2013. 416 с.
25. *Винберг Э.Б.* Курс алгебры. М.: МЦНМО, 2019. 592 с.
26. *Kaiser H.F.* // Educ. Psych. Measur. 1960. Vol. 20. N 1. P. 141. doi 10.1177/001316446002000116
27. *Cattell R.B.* // Multivariate Behav. Res. 1966. Vol. 1. N 2. P. 245. doi 10.1207/s15327906mbr0102_10
28. *Fisher R.A.* // Ann. Eugen. 1936. Vol. 7. N 2. P. 179. doi 10.1111/j.1469-1809.1936.tb02137.x
29. Халафян А.А. Современные статистические методы медицинских исследований. М.: ЛКИ, 2008. 320 с.
30. *Боровиков В.П.* Нейронные сети. Statistica Neural Networks. Методология и технологии современного анализа данных. М.: Горячая линия – Телеком, 2008. 392 с.
31. *Nocedal J., Wright S.J.* Numerical Optimization. Dordrecht: Springer, 2006. 683 p.
32. *Al-Baali M., Spedicato E., Maggioni F.* // Optimization Methods and Software. 2013. Vol. 29. N 5. P. 937. doi 10.1080/10556788.2013.856909
33. *Aggarwal C.C.* An Introduction to Neural Networks. In: Neural Networks and Deep Learningю New York: Springer, 2018. P. 1. doi 10.1007/978-3-319-94463-0_1752
34. *Izatt R.M., Bradshaw J.S., Nielsen S.A., Lamb J.D., Christensen J. J., Sen D.* // Chem. Rev. 1985. Vol. 85. N 4. P. 271. doi 10.1021/cr00068a003
35. *Filipek S., Wagner-Czauderna E., Kalinowski M.K.* // J. Coord. Chem. 1999. Vol. 48. N 2. P. 147. doi 10.1080/00958979908027962

Computer Analysis of Stability of Alkaline Metal Cation $M[222]^+$ Cryptates in Different Solvents

N. V. Bondarev*

V.N. Karazin Kharkiv National University, Kharkiv, 61022 Ukraine
**e-mail: bondarev_n@rambler.ru*

Received December 22, 2020; Revised December 22, 2020; accepted January 15, 2021

Computer analysis of the thermodynamic constants of complexation of cryptand [222] with alkali metal cations (cryptates $M[222]^+$, where M = Li, Na, K, Rb, Cs) in water and organic solvents such as methanol, ethanol, 1-propanol, acetonitrile, benzonitrile, acetone, *N*,*N*-dimethylformamide, *N*-methylpyrrolidone, nitrobenzene, nitromethane, 1,2-dichloroethane, and propylene carbonate at 298.15 K was performed. Exploratory (factorial, cluster, discriminant, canonical, decision tree), regression and neural network models of effects of the properties of solvents and cations on the cation cryptates stability were created. The neural network approximator MLP 4-7-1 and the classifiers of the stability constants of cryptates – the multilayer perceptron MLP 4-7-4 and the self-organizing Kohonen network SOFM 8-4 – were trained. Independent data on the stability constants of alkali metal cations cryptates demonstrate the predictive capabilities of the trained MLP 4-7-1 perceptron approximator.

Keywords: cryptand [222], complex formation constant, exploratory analysis, multiple linear regression, neural networks, modeling, forecasting