ЛЕКЦИЯ 13. ПОИСК НОВЫХ ТЕРМОЭЛЕКТРИЧЕСКИХ МАТЕРИАЛОВ. ВВЕДЕНИЕ В ХЕМОИНФОРМАТИКУ

В данной лекции рассмотрим методы хемоинформатики, узнаем, как с их помощью можно исследовать пространство материалов, в том числе термоэлектриков.

Термоэлектрики – это материалы, в которых за счет градиента температур возникает разность потенциалов. Они могут быть полупроводниками p- или n-типа. Оба типа нужны для создания термоэлектрических устройств (рис. 1).

Рисунок 1. Типы потенциалов

Данный эффект возникновения потенциала за счет градиента температуры называется эффектом Зеебека. Показатель эффективности:

$$zT = \frac{\alpha^2 \sigma T}{k_L + k_e}$$

Где α – коэффициент Зеебека;

σ – электропроводность;

T – температура;

k_L – решеточная теплопроводность;

k_e – электронная теплопроводность.

Множество материалов, принадлежащих к разным классам, проявляет этот эффект (рис. 2). Однако у одних материалов эффективность

преобразования намного выше, чем у других. Для экономической жизнеспособности материала необходим показатель эффективности zT>2.

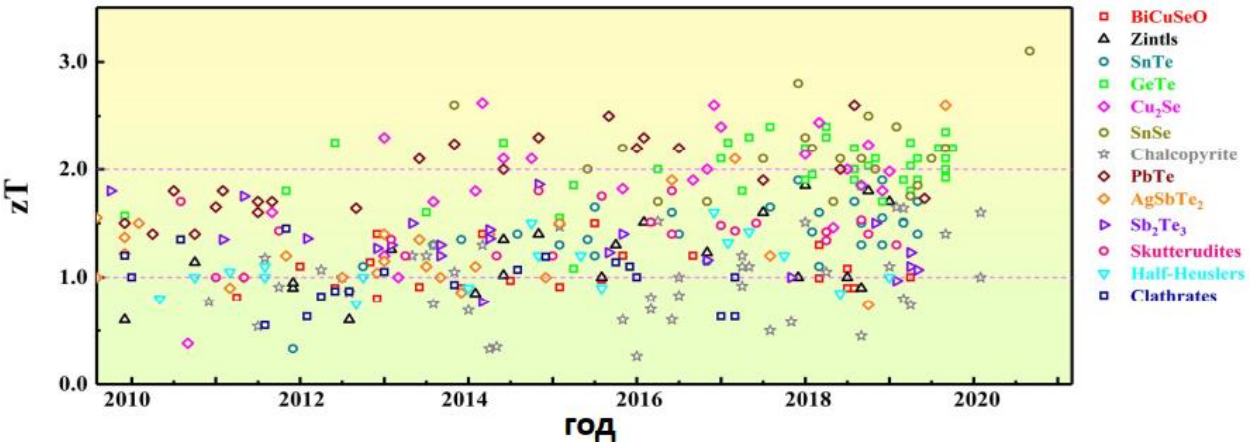


Рисунок 2. Данные по термоэлектрическим материалам

За последние несколько лет было открыто множество материалов с показателем эффективности более 2, однако многие материалы не являются практически пригодными из-за высокой стоимости и/или низкой стабильности. Проблема дизайна материалов заключается в сложном взаимодействие различных факторов, влияющих на величину zT.

Возникает вопрос: как можно предсказать новый материал? Существует два возможных способа дизайна материалов (рис. 3):

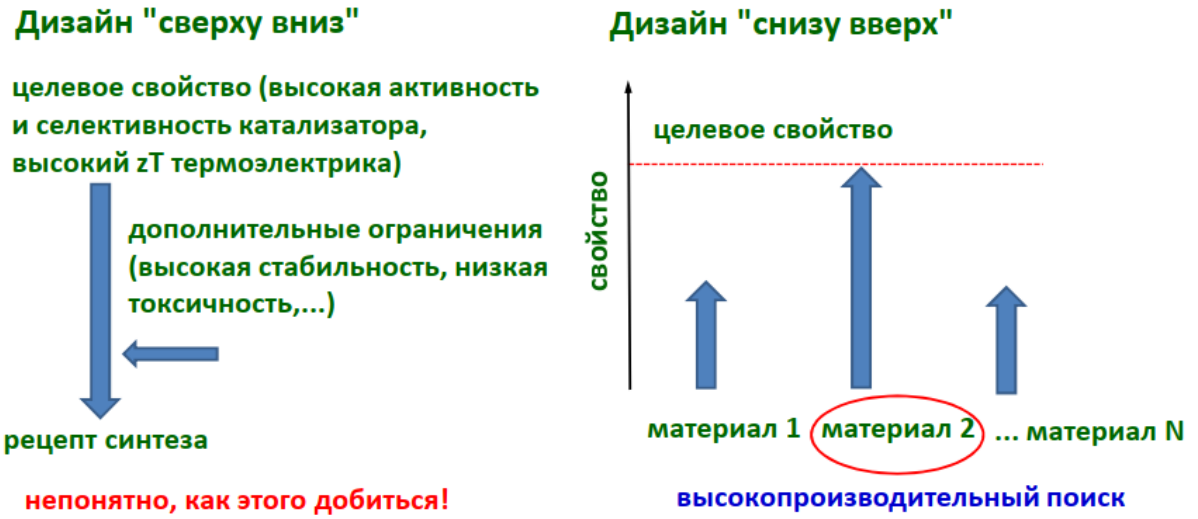


Рисунок 3. Сравнение способов дизайна материалов

– дизайн «сверху вниз». Начинаем с целевого свойства, например, высокого показателя эффективности термоэлектрика или высокой активности и селективности катализатора. Существуют дополнительные ограничения: низкая токсичность, низкая стоимость, высокая стабильность. На основе этих входных данных предсказывается рецепт синтеза материала – идеальное решение задачи. Однако, в общем, непонятно, как этого добиться. Такой подход разрабатывается для узкого класса материалов, например, для органических материалов.

– дизайн «снизу вверх». Используется чаще всего. Идея состоит в том, что мы рассчитываем или предсказываем свойства, которое необходимо оптимизировать для большого числа материалов, затем выбираем наиболее оптимальный для нашего приложения. Такой подход называется высокопроизводительный поиск.

Ключевая проблема теоретического дизайна материалов (рис. 4) заключается в сложности систем.

$$i\frac{\partial\Psi(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n,\boldsymbol{R}_1,\boldsymbol{R}_2,\ldots,\boldsymbol{R}_N,t)}{\partial t}=\widehat{H}(t)\Psi(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n,\boldsymbol{R}_1,\boldsymbol{R}_2,\ldots,\boldsymbol{R}_N,t)$$

1) **Многочастичная задача ($3(n+N)$- размерная)**

2) **Многомасштабная задача (десятки порядков во времени и пространстве)**

Рисунок 4. Ключевая проблема теоретического дизайна материалов – сложность системы

Первая причина заключается в многочастичности задачи (3(n + N) – размерная). Волновая функция зависит от огромного числа переменных, это математически сложная задача.

Другая причина – масштабность задачи - необходимо учесть различные химические процессы и многое др. (десятки порядков во времени и пространстве). Тем не менее, есть надежда, что проблему сложности можно решить пошагово. Например (рис. 5), можно определить какие-либо свойства материала, которые могут быть вычислены или измерены легко, и на их основе предсказать значение целевого свойства.

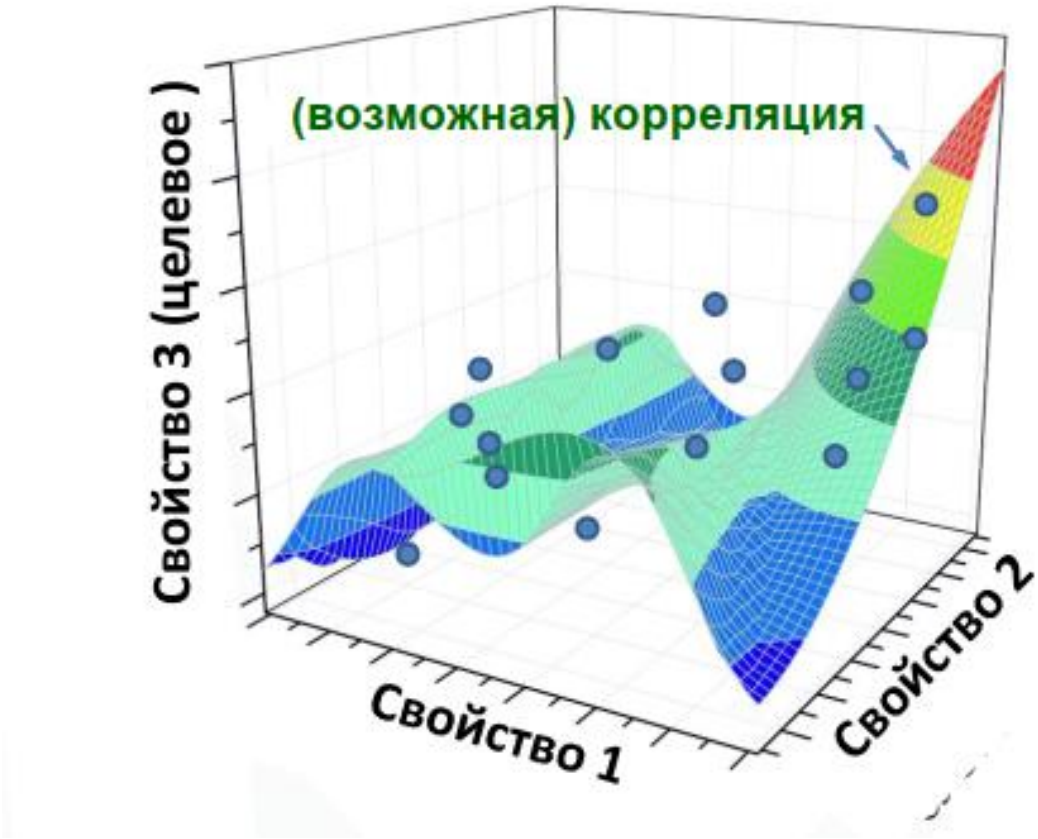


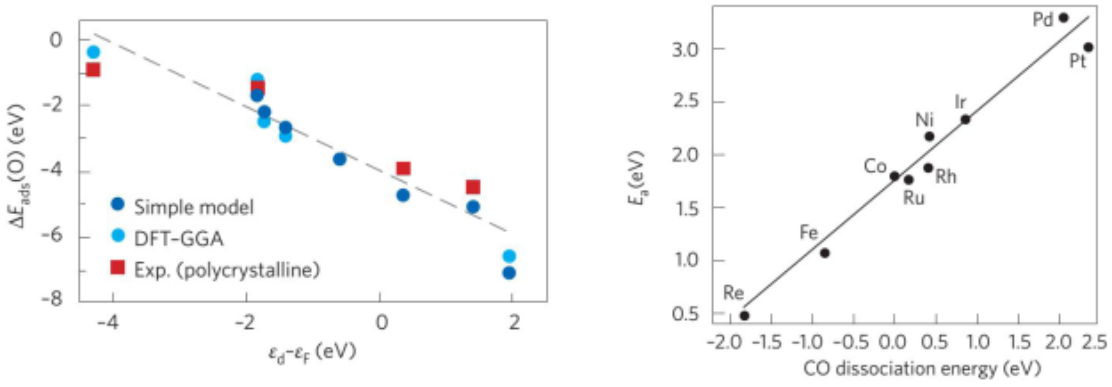
Рисунок 5. Поверхность потенциальной энергии в зависимости от свойств материалов

Например, рассчитать показатель эффективности термоэлектрика сложно, но возможно определить кристаллическую и электронную структуру материала, и на их основе предсказать целевое свойство. То есть необходимо найти поверхность потенциальной энергии, где вместо координат – свойства материала, а вместо энергии – целевое свойство.

Одним из ярких примеров подобного подхода является использование дескрипторов при поиске материалов для гетерогенного катализа. Дескрипторы являются свойствами, которые проще рассчитать, чем целевые.

Напомним, что катализ – это процесс ускорения химических реакций, катализаторы – материалы, которые ускоряют химические реакции, при этом сами сохраняются. В гетерогенном катализе на твердой поверхности происходят химические реакции. Сложность задачи заключается в том, что химические реакции на поверхности изменяют непосредственно поверхность, и в реальных условиях данные процессы неизвестны. Для решения данной задачи необходимы упрощения, величины, с помощью которых мы сможем предсказать каталитическую активность.

Яркий пример в этой области продемонстрировали J.K. Nørskov et al. в 2009 г. (рис. 6), показав, что существует корреляция между энергией адсорбции атома или молекулы на поверхностях различных металлов и центрами d-зоны.



Простая физическая модель (Ньюнс-Андерсон) мотивирует дескриптор центра d-зоны

Найти дескриптор из ДАННЫХ!

J. K. Nørskov, T. Bligaard, J. Rossmeisl and C. H. Christensen, Nature Chemistry **1**, 37 (2009)

Рисунок 6. Корреляция между энергией адсорбции атома или молекулы на поверхностях различных металлов и центрами d-зон

Для расчета энергии адсорбции необходимо взять большую систему атомов - суперячейку - и помещать в нее атомы в разных точках, проводить релаксацию системы - расчет занимает много времени и вычислительных ресурсов, а для расчета d-зоны необходима только чистая поверхность. Существует хорошая линейная корреляция между центром d-зоны и энергией адсорбции - это пример дескриптора. Подобная корреляция существует между энергией активации реакции (определяется барьером реакции - его расчет занимает много времени и вычислительных ресурсов) и энергией диссоциации молекулы, которую рассчитать гораздо проще.

Что делать, если мы не знаем модели, или нам нужна более точная модель? Идея состоит в том, чтобы найти дескрипторы из имеющихся данных. Общая схема поиска представлена на рис. 7.

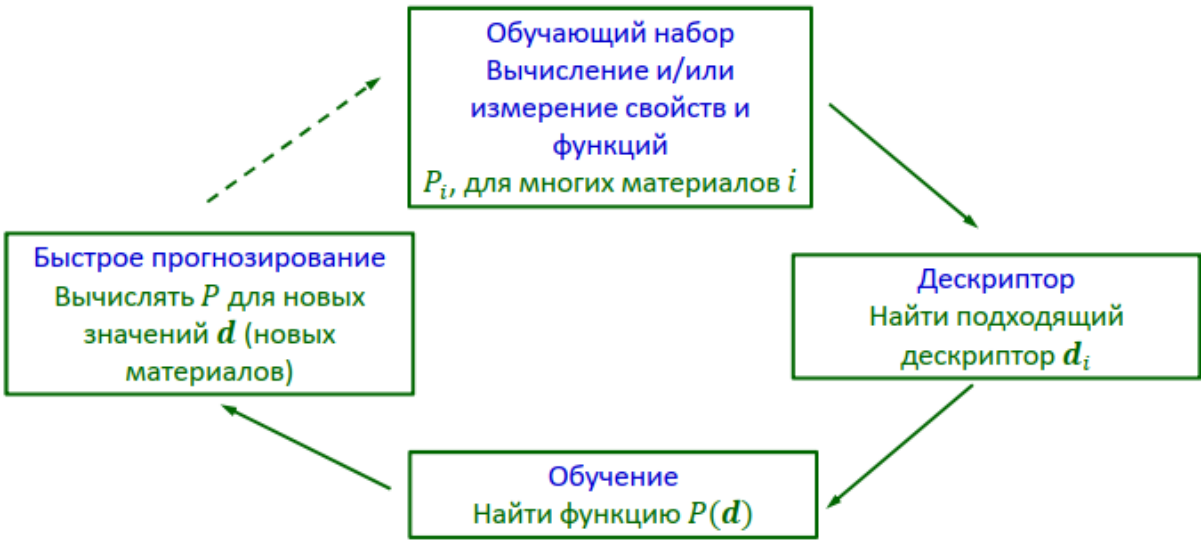


Рисунок 7. Поиск модели на основе анализа данных

Существует обучающий набор данных, то есть значения дескрипторов или основных характеристик материалов и вычисленные целевые свойства для некоторых из них. Из данного набора определяем подходящий дескриптор, затем находим функцию зависимости свойства от дескриптора и прогнозируем свойства для других материалов. Затем снова проводим расчет или

эксперимент для отобранных материалов и включаем их в обучающие данные, улучшая модель. Данный цикл называется циклом активного обучения.

Сейчас для поиска дескрипторов популярно применять методы искусственного интеллекта, включая машинное обучение. Множество подобных методов запрограммированы в scikit-learn, в том числе нейронные сети, байесовский вывод, кластеризация, гребневая регрессия с ядром, символическая регрессия, деревья решений, интеллектуальный анализ данных. Соответственно, существует целый набор методов решения задачи со своими преимуществами и недостатками. Наиболее классическими методами являются регрессия и нейронные сети.

Авторы данного учебного курса разрабатывают методы символьной регрессии на основе сжатого зондирования и метода обнаружения подгрупп. Преимущество данных методов – возможность работы с небольшими объемами данных. Например, можно использовать экспериментальные данные для тренировочного набора (в эксперименте мы не можем охарактеризовать тысячи материалов, а нейронные сети требуют большой объем данных). Также, преимуществом является возможность физической интерпретации результатов.

Дескриптор d_i должен уникально характеризовать материал i, а также элементарные процессы, имеющие отношение к его свойствам. Определение дескриптора не должно включать в себя столь же интенсивные расчеты, как те, которые необходимы для оценки прогнозируемого свойства.

Рассмотрим гребневую регрессию с ядром (рис. 8).

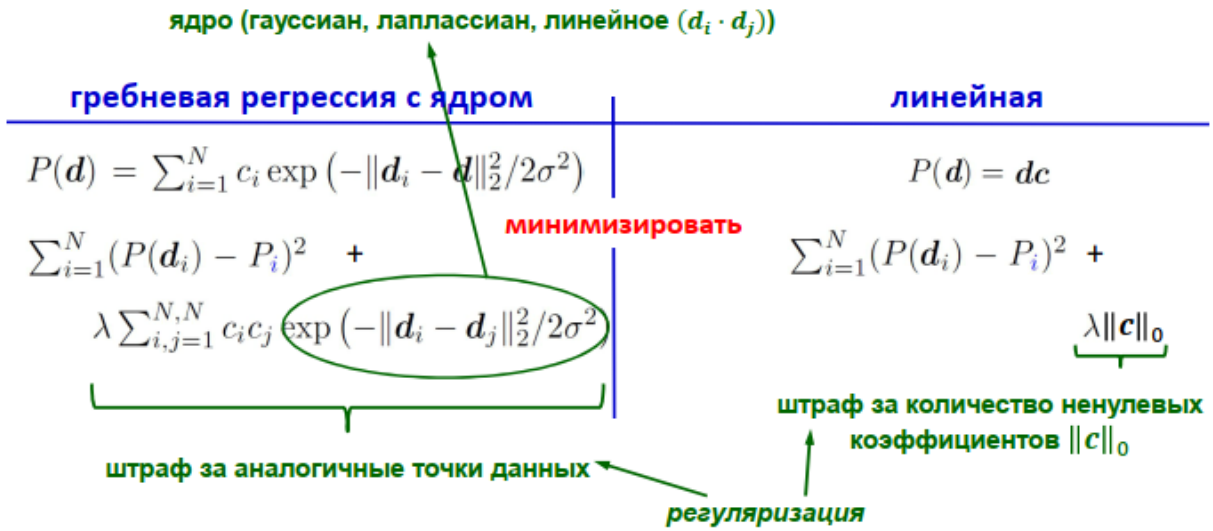


Рисунок 8. Модель гребневой регрессии с ядром в сравнении с линейной регрессией

d_i – это дескрипторы для тренировочных данных, для материалов с известным дескриптором и целевым свойством,

d – это дескриптор для неизвестного материала,

C_i – коэффициенты,

Pi – значение целевого свойства для тренировочного набора данных (уже посчитали или измерили).

В гребневой регрессии целевое свойство, например, показатель эффективности термоэлектрика, выражается как функция дескрипторов - линейная комбинация неких функций, например, Гауссова экспонента. Гребневая регрессия с Гауссовым ядром наиболее популярна. Она очень часто используется в изучении материалов.

По сути это похоже на разложение по базису нашей неизвестной функции целевого свойства. Существует некая функция и базисный набор, который представляет собой гауссианы, центрированные на d_i точках. Задача состоит в том, чтобы найти коэффициенты C_i для минимизации ошибки предсказаний. Например, можно использовать в качестве метрики декартово расстояние в пространстве дескрипторов.

В линейной регрессии непосредственно дескриптор выступает в виде базисного набора (базиса). Свойства представляются как линейная комбинация компонентов дескриптора с некими коэффициентами. Задача – минимизировать ошибку.

Задача минимизации определена - число уравнений равно числу переменных только в случае, если число коэффициентов совпадает с числом тренировочных данных. Если коэффициентов (параметров модели) мало (рис. 9), то модель будет намного проще настоящей корреляции данных, соответственно, их будет невозможно воспроизвести - происходит Underfitting – недостаточно точная подгонка, характеризующаяся большой ошибкой предсказания новых материалов. В случае, когда у нас очень много коэффициентов, можно очень точно подогнать модель под тренировочные данные. В таком случае происходит Overfitting. Ошибка подгонки будет малой, однако ошибка предсказания может быть еще больше, чем в случае Underfitting.

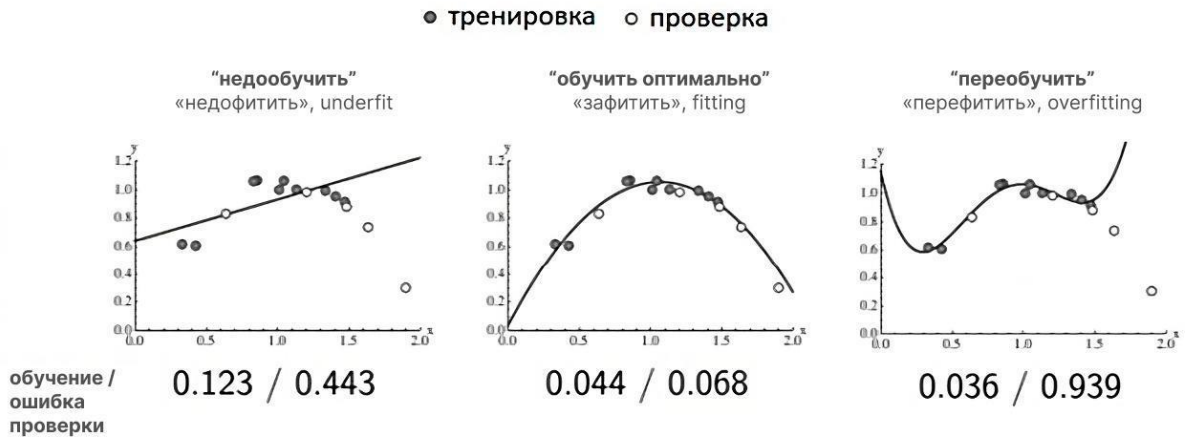


Рисунок 9. Обучение модели: Underfitting, Fitting, Overfitting

Нам необходимо определить промежуточный вариант – Fitting, характеризующийся оптимальным количеством параметров. Он даст хорошее описание данных и высокую точность предсказаний. Данный вариант

достигается отбором коэффициентов (ненулевых) путем регуляризации. Добавляется дополнительный член: $\lambda f(c)$, затем функция минимизируется по отношению к коэффициентам.

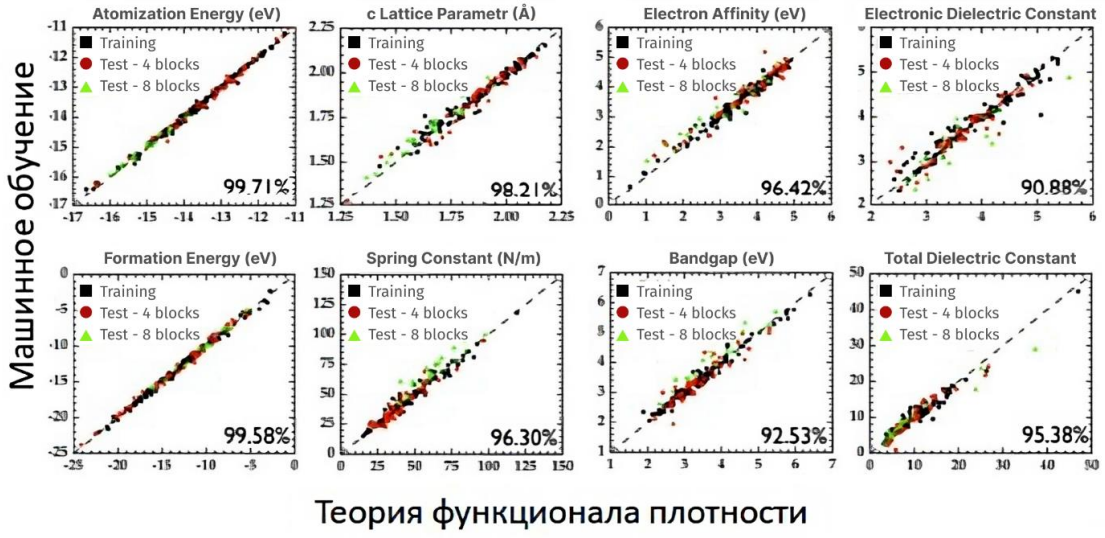
В случае гребневой регрессии с ядром – это дополнительное слагаемое представляет собой более сложную функцию:

$$\lambda \sum_{i,j=1}^{N.N} \quad c_i c_j exp\,(-\|d_i - d_j\|_2^2 / 2\sigma^2)$$

Смысл данного слагаемого следующий: когда дескриптор d_i и дескриптор d_j близки друг к другу, модель без регуляризации использует их как разные, что не дает информации, только увеличивает число коэффициентов и приводит к переопределению модели. Дополнительный член уменьшает роль близлежащих данных, тем самым сокращая ошибку.

В случае линейной регрессии, когда дескриптор используется в качестве базиса, можно наложить штраф за количество ненулевых коэффициентов, чтобы уменьшить число базисных функций.

В качестве примера приведем работу Ramprasad и соавторов, которые исследовали полимеры и их диэлектрические свойства (рис. 10).



Pilania, Wang, ..., and Ramprasad, Scientific Reports 3, 2810 (2013). DOI: 10.1038/srep02810

Рисунок 10. Исследование полимеров и их диэлектрических свойств методом гребневой регрессии

Полимеры используются в качестве диэлектриков в конденсаторах. Важно найти диэлектрики, характеризующиеся большой диэлектрической постоянной и устойчивостью к пробиванию при высоких напряжениях. Была использована гребневая регрессия. Данные включали 175 линейных 4-блоковых периодических полимеров, состоящих из: CH_2, SiF_2, $SiCl_2$, GeF_2, $GeCl_2$, SnF_2, $SnCl_2$. Дескриптор представлял собой набор блоков типа *i*, пар *ij*, триплетов *ijk* - использовался многоразмерный дескриптор.

Результаты показаны на рис. 10. По оси Y отложено предсказание, сделанное с помощью машинного обучения, по оси X – с помощью теории функционала плотности. Модель работает хорошо, но для одних свойств лучше, чем для других хуже. Например, энергия атомизации и энергия образования описаны хорошо, а для диэлектрической постоянной существуют отклонения. Однако есть проблема, заключающаяся в утере физического понимания (интерпретируемости) этой модели.

Поэтому была поставлена задача найти модель, в которой размерность дескриптора была бы как можно ниже, в идеале один параметр – центр d-зоны, или два-три (в зависимости от точности, которая необходима для предсказания). Идея заключается в выборе физически мотивированного базисного набора.

Классическим примером является определение кристаллической структуры (рис. 11).

Рисунок 11. Классификации структур «Цинковая смесь/Вюрцит (ZB/W) или каменная соль (RS)?»

Необходимо найти дескриптор для классификации структур -Цинковая смесь/Вюрцит (ZB/W) или каменная соль (RS). Необходимо предсказать, в какой из этих кристаллических структур будет находиться бинарный материал типа AB. Сложность заключается в малых энергетических различиях между структурами.

Как известно, все свойства через уравнение Шредингера определяются зарядами ядер атомов, например, A и B. Но если создать такую карту, как на рис. 12, показывающую разности энергий между структурами каменной соли и цинковой смеси для различных материалов, мы увидим сложную зависимость, которую будет трудно интерпретировать.

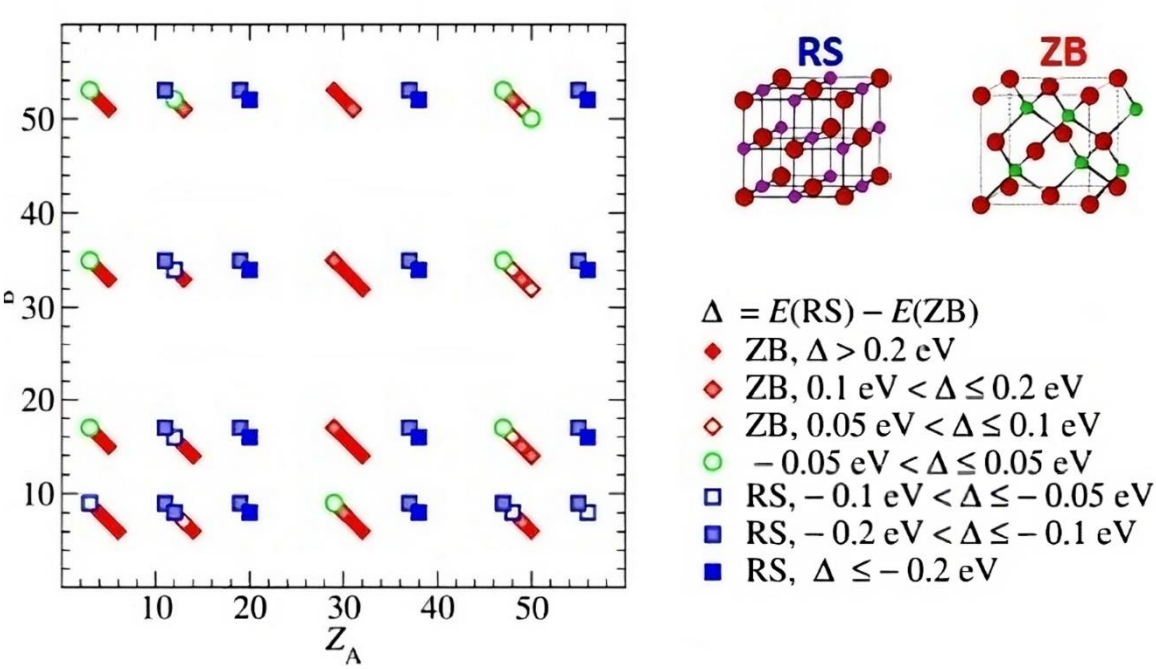


Рисунок 12. Карта, показывающая разности энергий между структурами каменной соли и цинковой смеси для различных материалов

Мы не увидим распределения по зонам как на рис. 13, в котором существовали бы величины d_1 и d_2, и все материалы со структурой каменной соли находились бы в одной (синей) зоне, со структурой цинковой смеси – в другой (красной), а вырожденные - в третьей (зеленой).

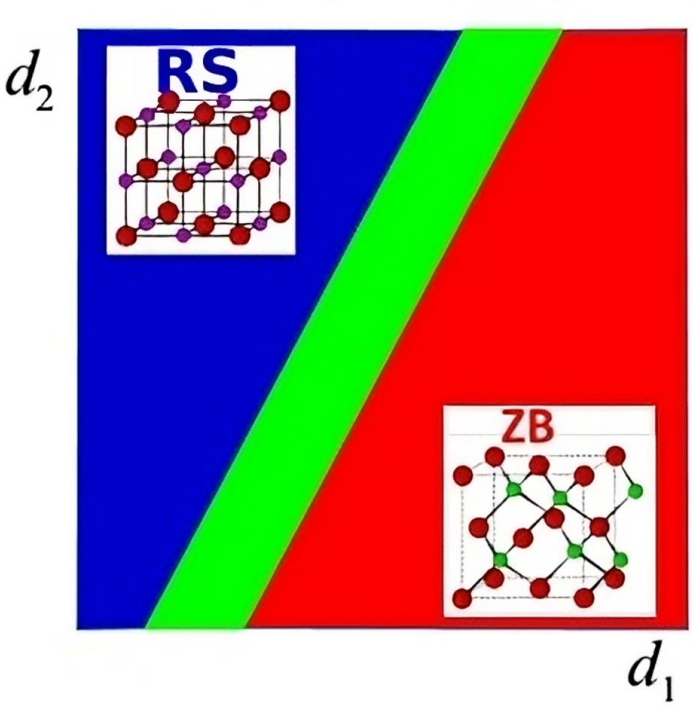


Рисунок 13. Распределение материалов по зонам в зависимости от их структуры в пространстве дескрипторов d_1 и d_2

В работах J.A. van Vechten и J. C. Phillips, были предложены величины E_h и C, которые относятся к ширине запрещенной зоны и могут быть определены спектроскопически, т.е. экспериментально. На основе их получается карта (рис. 14), на которой линией разделены материалы со структурой каменной соли (верхняя часть) и со структурой цинковой смеси (нижняя часть), а также материалы, которые нечетко определяются по отношению к группам.

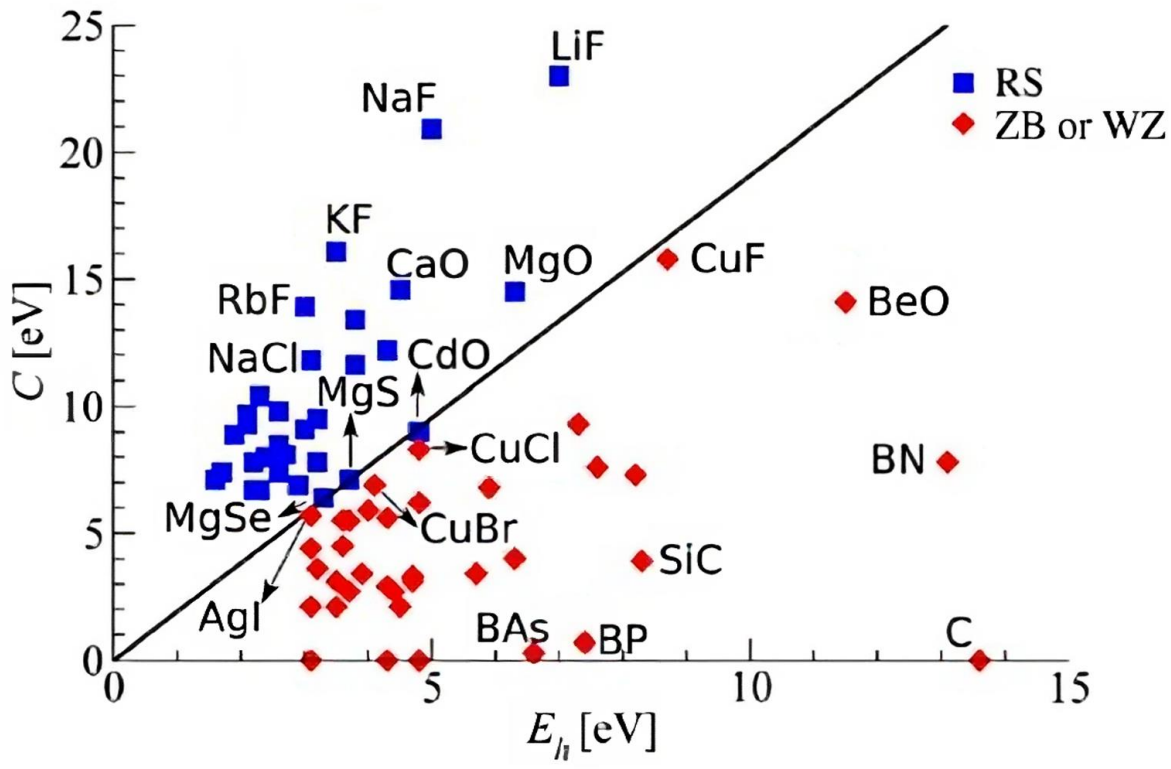


Рисунок 14. Распределение материалов по зонам в зависимости от их структуры на основе спектроскопических данных

С точки зрения экспериментальной физики – это интересный опыт. Вместо затратной рентгеновской спектроскопии для определения кристаллической структуры были измерены спектроскопические свойства и предсказаны кристаллические структуры материалов. Однако, с теоретической точки зрения, в этом нет никакой ценности, так как для определения данных величин необходимо посчитать зонную структуру материала, которая и даст нам значения энергий материалов - по затратам это невыгодно.

Возникает вопрос: «Можно ли создать карту на основе расчетов более простых свойств, чем свойства кристаллов?». Можно. С помощью методов искусственного интеллекта возможно определить такие свойства атомов A и B, как: потенциалы ионизации, сродство к электрону, энергии орбиталей, свойства димеров - эти свойства (рис. 15) называются основными характеристиками или основными признаками материалов.

Изолированные атомы

ID	Описание	Обозначение	#
A1	потенциал ионизации (IP) и электроотрицательность (EA)	IP(A) EA(A) IP(B) EA(B) [1]	4
A2	высший заселенный (H) и низший незаселенный (L) уровни Кона - Шэма (KS)	H(A) L(A) H(B) L(B)	4
A3	радиусы, на которых радиальные плотности вероятности валентных s, p и d орбиталей достигают максимумов	r_s(A) r_p(A) r_d(A) r_s(B) r_p(B) r_d(B)	6

Изолированные димеры

ID	Описание	Обозначение	#
A4	энергия связи E_b	E_b(AA) E_b(BB) E_b(AB)	3
A5	HOMO-LUMO KS щели	HL(AA) HL(BB) HL(AB)	3
A6	равновесное расстояние	d(AA) d(BB) d(AB)	3

Рисунок 15. Основные характеристики (признаки) материалов

Как же найти лучшую модель для целевого свойства? Существует метод символьной регрессии. Он был реализован в программе Eureqa. Мы берем основные признаки, комбинируем их с математическими операторами (рис. 16), создаем множество комбинаций и с помощью эволюционного алгоритма ищем формулу, описывающую целевое свойство наилучшим образом. Идея похожа на идею программы «USPEX», в которой для поиска кристаллической структуры сначала генерируется множество различных атомных конфигураций и затем, с помощью эволюционного алгоритма, выбирается наилучшая из них.

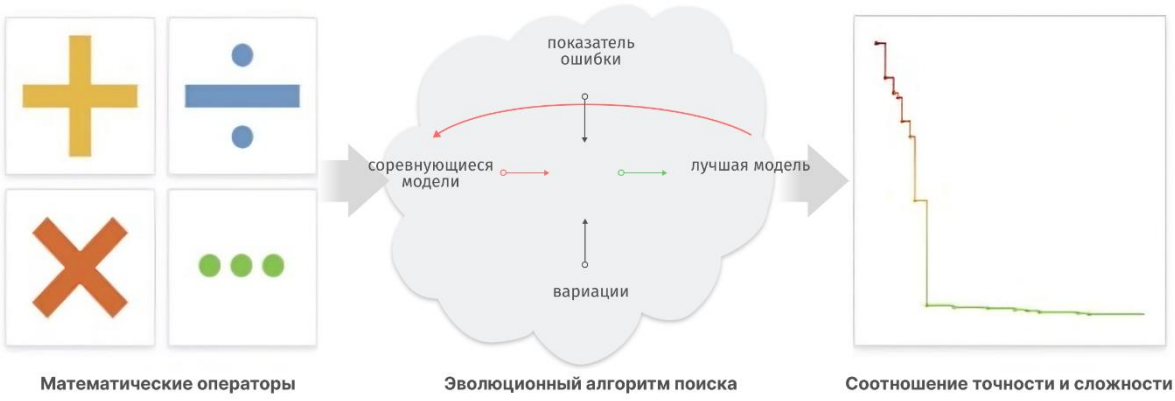


Рисунок 16. Метод символьной регрессии, реализованный в программе Eureqa

Другой подход (рис. 17) начинается аналогично, то есть с генерации различных комбинаций с помощью математических операторов из основных признаков - в результате получается более десяти тысяч нелинейных комбинаций.

Изолированные атомы

ID	Описание	Обозначение	#
A1	потенциал ионизации (IP) и электроотрицательность (EA)	IP(A) EA(A) IP(B) EA(B) [1]	4
A2	высший заселенный (H) и низший незаселенный (L) уровни Кона - Шэма (KS)	H(A) L(A) H(B) L(B)	4
A3	радиусы, на которых радиальные плотности вероятности валентных s, p и d орбиталей достигают максимумов	r_s(A) r_p(A) r_d(A) r_s(B) r_p(B) r_d(B)	6

Изолированные димеры

ID	Описание	Обозначение	#
A4	энергия связи E_b	E_b(AA) E_b(BB) E_b(AB)	3
A5	HOMO-LUMO KS щели	HL(AA) HL(BB) HL(AB)	3
A6	равновесное расстояние	d(AA) d(BB) d(AB)	3

ID	Описание	Формула	#
B1	модули разностей и сумм A1	$\lvert IP(A) \pm IP(B) \rvert$	12
B2	модули разностей и сумм A2	$\lvert L(B) \pm H(A) \rvert$	12
B3	модули разностей и сумм A3	$\lvert r_p(A) \pm r_s(A) \rvert$	30
C3	квадраты A3 и B3 (только сумм)	$r_s(A)^2$,$(r_p(A) + r_s(A))^2$	21
D3	экспоненты A3 и B3 (только сумм)	$\exp(r_s(A))$, $\exp(r_p(A) \pm r_s(A))$	21
E3	экспоненты квадратов A3 и B3 (только сумм)	$\exp(r_s(A)^2)$, $\exp(r_p(A) \pm r_s(A)^2)$	21

Мы начинаем с 23 основных признаков и создаем > 10 000 нелинейных комбинаций

Рисунок 17. Основные признаки материалов, используемые для генерации нелинейных комбинаций

Они используются как кандидаты в дескрипторы. Величина свойства, представляющая собой разность энергий между двумя различными кристаллическими структурами для каждого материала в тренировочном наборе - это функция в пространстве материалов. А кандидаты в дескрипторы представляют собой базисные функции в пространстве материалов. Поэтому

свойства – это линейная комбинация дескрипторов с какими-то коэффициентами.

Как найти эти коэффициенты? Необходимо использовать регуляризацию, чтобы избежать переопределений или недоопределений системы и снизить ее сложность. Мы хотели бы перепробовать все комбинации (математическая постановка задачи представлена на рис. 18), но на практике это невозможно.

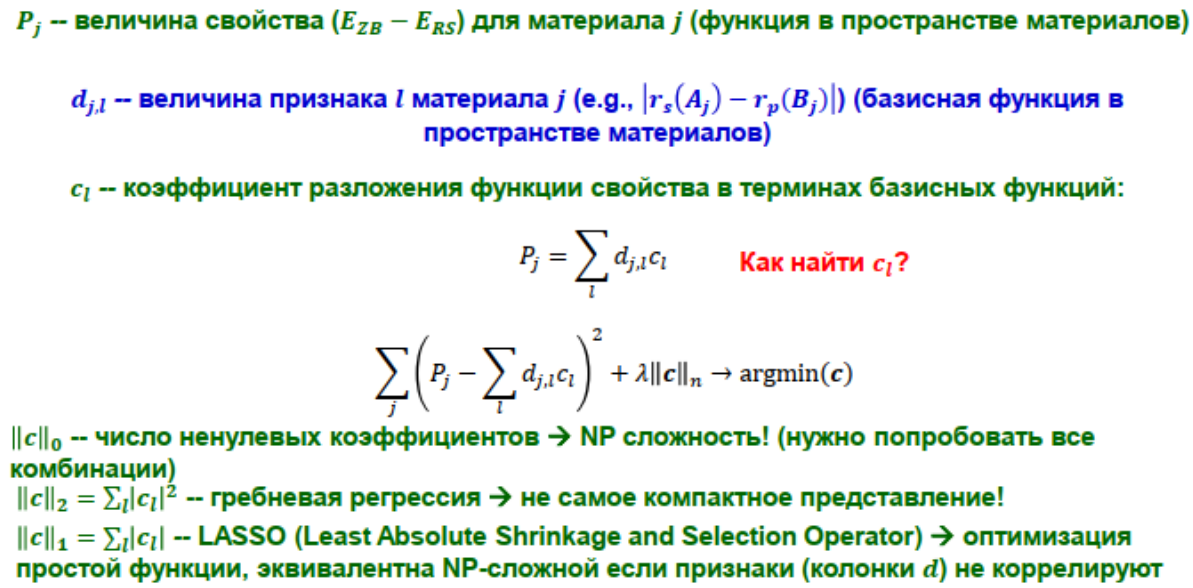


Рисунок 18. Математическая постановка задачи поиска коэффициентов

Поэтому используется гребневая регрессия, где в качестве штрафа за ненулевые коэффициенты применяется сумма квадратов коэффициентов. Однако это не самый компактный способ представления, ведь остается много ненулевых коэффициентов.

Для решения данной проблемы применяется метод LASSO (Least Absolute Shrinkage and Selection Operator), который незначительно отличается от гребневой регрессии - вместо суммы квадратов стоит сумма модулей. Существует теорема, которая доказывает, что оптимизация функции эквивалентна полному перебору (при условии, что признаки (дескрипторы) не

коррелируют друг с другом). Такой метод используется во многих областях, в основном, в анализе сигналов.

На рис. 19 представлен пример сжатия изображения. Картинка была сжата почти в 100 раз. Идея аналогична той, которую мы хотим использовать для дизайна материалов: изображение разлагается по базису (вейвлеты – локализованный базис). LASSO используется для выбора наиболее важных базисных функций. Параметры этих функций сохраняются. В результате получается сжатое изображение - это принцип JPEG. Данный подход называется сжатое или сжимающее зондирование.

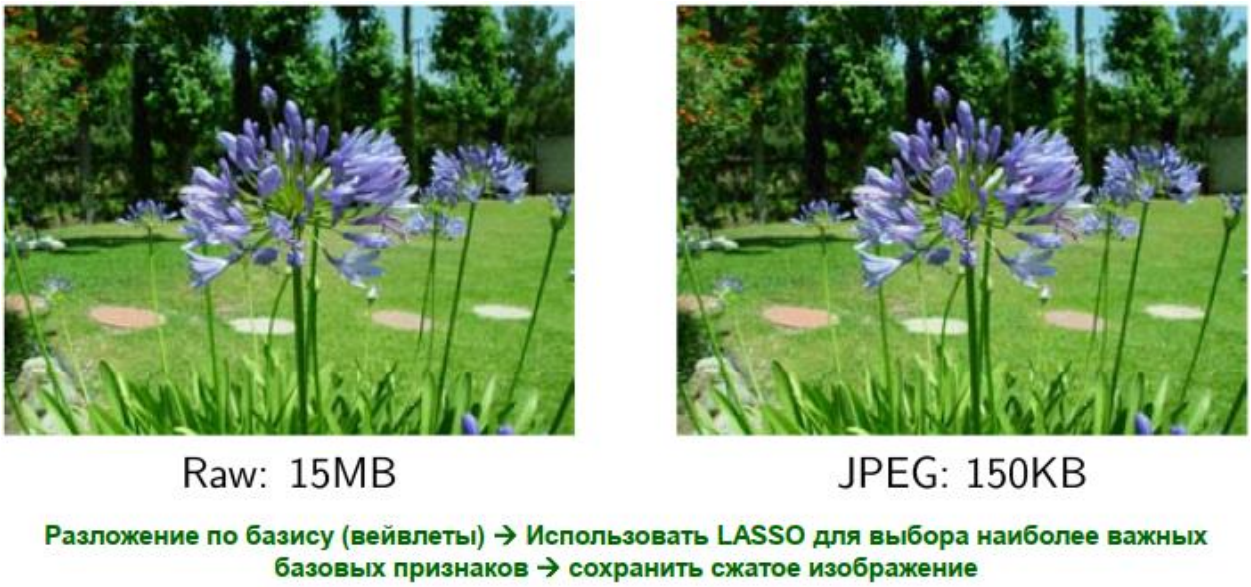


Рисунок 19. Иллюстрация сжатого (сжимающего) зондирования

На рис. 20 представлено пространство коэффициентов. Число неизвестных и число дескрипторов равно 2, значение дескрипторов составляет 10 и 7 при коэффициентах *x* и *y* - C_1 и C_2 (вдоль оранжевой линии свойство имеет одно и то же значение, например, 20). Минимум задачи (если n = 2) будет, когда окружность вокруг нуля (линия, где сумма квадратов коэффициентов постоянна) при определенном значении λ достигает точки с точным решением, соответствующее двум ненулевым коэффициентам.

P_j -- величина свойства ($E_{ZB} - E_{RS}$) для материала j (функция в пространстве материалов)

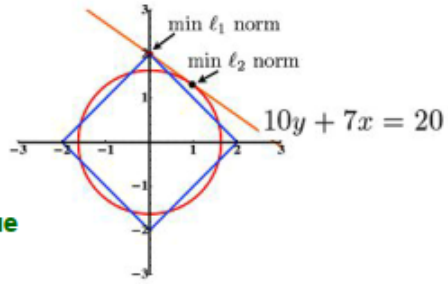
$d_{j,l}$ -- величина признака l материала j (e.g., $|r_s(A_j) - r_p(B_j)|$) (базисная функция в пространстве материалов)

c_l -- коэффициент разложения функции свойства в терминах базисных функций:

$$P_j = \sum_l d_{j,l} c_l$$

Как найти c_l?

$$\sum_j \left(P_j - \sum_l d_{j,l} c_l \right)^2 + \lambda \|\boldsymbol{c}\|_n \rightarrow \operatorname{argmin}(\boldsymbol{c})$$

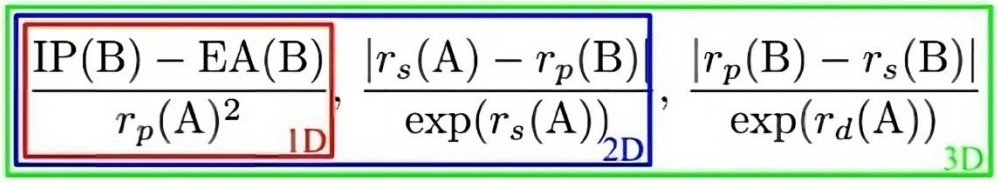


$\|c\|_1 = \sum_l |c_l|$ -- LASSO (Least Absolute Shrinkage and Selection Operator) → оптимизация простой функции, эквивалентна NP-сложной если признаки (колонки $\boldsymbol{d}$) не коррелируют (отсутствие линейной зависимости в базисном наборе)

Рисунок 20. Математическая постановка задачи

Когда n = 1, что в двумерном пространстве соответствует синему квадрату, а в многомерном – гиперкубу, оптимальное решение возникает, когда этот куб касается линии только в одной точке. В этом заключается работа зондирования – анализ пространства и поиск решения с минимальным числом ненулевых коэффициентов.

Итак, мы применяем метод сжатого зондирования LASSO к проблеме предсказания разностей энергий различных кристаллических структур. Они выражаются следующим образом (рис. 21):



$$\Delta E = 0.117\frac{\mathrm{EA(B)}-\mathrm{IP(B)}}{r_p(\mathrm{A})^2} - 0.342 \quad \text{1D}$$

$$\Delta E = 0.113\frac{\mathrm{EA(B)}-\mathrm{IP(B)}}{r_p(\mathrm{A})^2} + 1.542\frac{|r_s(\mathrm{A})-r_p(\mathrm{B})|}{\exp(r_s(\mathrm{A}))} - 0.137 \quad \text{2D}$$

$$\Delta E = 0.108\frac{\mathrm{EA(B)}-\mathrm{IP(B)}}{r_p(\mathrm{A})^2} + 1.790\frac{|r_s(\mathrm{A})-r_p(\mathrm{B})|}{\exp(r_s(\mathrm{A}))} + \quad \text{3D}$$

$$+ 3.766\frac{|r_p(\mathrm{B})-r_s(\mathrm{B})|}{\exp(r_d(\mathrm{A}))} - 0.0267$$

Рисунок 21. Разности энергий различных кристаллических структур в одно-, дву- и трехмерном пространстве

Согласно полученным формулам, дескрипторы высшей размерности включают в себя те же самые компоненты, что и дескрипторы более низкой размерности. Но это не обязательно так.

Для двумерного дескриптора была построена карта. Как видно из рис. 22, на его основе получилось разделить материалы с различной кристаллической структурой. Символами представлены разные диапазоны разностей энергий. Область зеленых точек, где разность очень мала между двумя кристаллическими структурами, довольно размыта. Однако в целом модель работает.

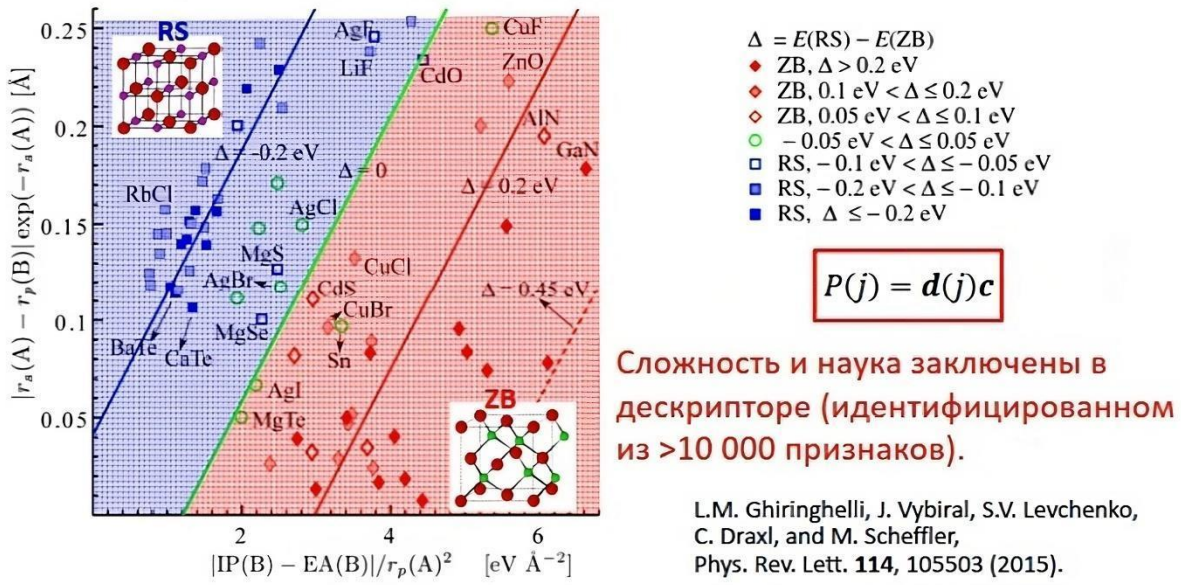


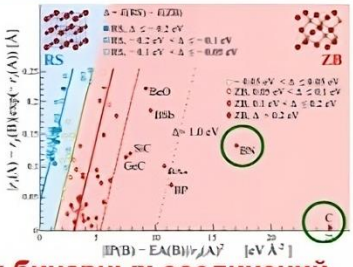
Рисунок 22. Карта структур, построенная на основе 2D - дескриптора

Насколько же хороша предсказательная сила такой модели? Мы убрали из набора данных нитрит бора и углерод, так как данные материалы однозначно предпочитают структуру цинковой смеси, и построили модель без этих материалов (тренировочные данные не содержали эти элементы). На основе вновь полученной модели мы смогли предсказать, что они предпочитают структуру цинковой смеси (даже порядок разности энергий сохранится). Даже если мы исключим все материалы, содержащие углерод, мы сможем предсказать их свойства, более того, в большинстве случаев сохранится даже порядок разностей энергий (рис. 23).

Если бы мы не знали об алмазе... мы бы его предсказали!

Когда из обучения исключены и углеродный алмаз, и BN:

	ΔE(LDA)	ΔE(predicted)
C	-2.64 eV	-1.44 eV
BN	-1.71 eV	-1.37 eV



Если бы мы не знали никаких углеродсодержащих бинарных соединений ... мы бы предсказали химию углерода (исходя из свойств атомов)

Когда все C-содержащие бинарные соединения (C, SiC, GeC, SnC) исключены из обучения, т.е. информация о C не включена явно в модель:

	ΔE(LDA)	ΔE(predicted)
C	-2.64 eV	-1.37 eV
SiC	-0.67 eV	-0.48 eV
GeC	-0.81 eV	-0.46 eV
SnC	-0.45 eV	-0.23 eV

Рисунок 23. Предсказательная сила модели

В этом и заключается идея хемоинформатики - в предсказании свойств материалов, даже не включенных в модель - в исследовании всего пространства материалов на основе небольшого объема исходных данных.

Проанализируем более количественно ошибку предсказаний. Возьмем в качестве дескрипторов заряды атомов, так как известно существование однозначной связи между зарядами и свойствами материалов (разностью энергий). Как видно из рисунка 24, гребневая регрессия справляется с подгонкой тренировочных данных, ошибки Za, Zb очень малы.

Средняя абсолютная погрешность (MAE), and максимальная абсолютная погрешность (MaxAE), в эВ, (первые две строки) и для перекрестной проверки путем 10% выборки (CV), усредненной для более 150 случайных выборов тренировочного набора (последние две строки). Для (Z_A*, Z_B*), каждый атом идентифицируется строкой из трех случайных чисел.

Descriptor	Z_A, Z_B	Z_A*, Z_B*	1D	2D	3D	5D
MAE	$1*10^{-4}$	$3*10^{-3}$	0.12	0.08	0.07	0.05
MaxAE	$8*10^{-4}$	0.03	0.32	0.32	0.24	0.20
MAE, CV	0.13	0.14	0.12	0.09	0.07	0.05
MaxAE, CV	0.43	0.42	0.27	0.18	0.16	0.12

Gaussian-kernel ridge regression

LASSO

Рисунок 24. Предсказательная сила модели

Существует перекрестный способ проверки модели. Алгоритм ее следующий: берется набор данных, из него убирается часть и определяется модель (по оставшимся данным), далее предсказываются свойства тех материалов, что убрали из первичного набора. Эта операция повторяется несколько раз для разных случайных выборок.

В рассматриваемом примере для перекрестной проверки выборка составляла 10% от тренировочных данных. Ошибка предсказаний усреднялась более чем для 150 случайных наборов тренировочных данных. В качестве дескрипторов использовались заряды A и B. Как видим, наблюдается хорошая точность подгонки, неплохая точность предсказания, однако максимальная ошибка – большая.

LASSO дает большие ошибки подгонки, но они могут быть систематически уменьшены с помощью увеличения размерности компонентов дескриптора. Также уменьшается максимальная ошибка – модель начинает “улавливать” какие-то физические закономерности.

В целом, существует связь между свойствами материала и его целевым свойством. Она очень сложна. Дескрипторы не определяют напрямую целевое свойство - сложная причинно-следственная связь (рис. 25).

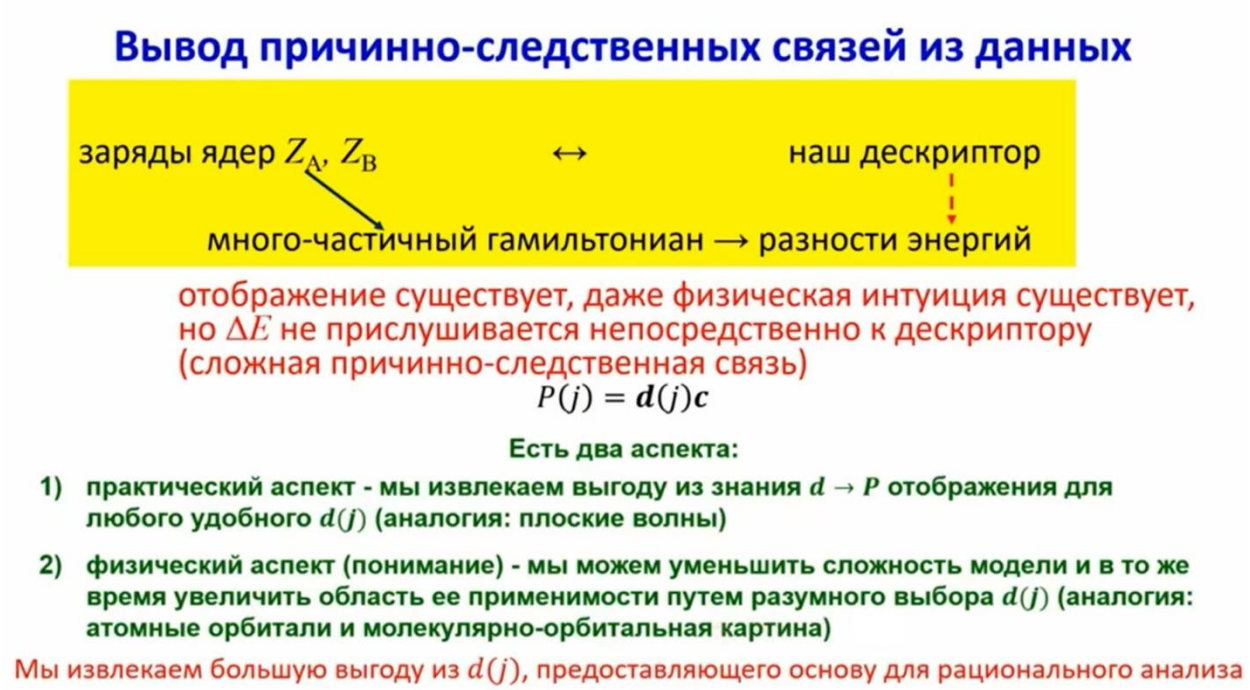


Рисунок 25. Вывод причинно-следственных связей из данных

Существует два аспекта:

- практический аспект – извлекаем выгоду из знания соотношения между дескриптором и свойством при условии, что дескриптор легко посчитать или измерить;
- физический аспект – можно уменьшить сложность модели и в то же время увеличить область ее применения путем разумного выбора дескрипторов.

Мы извлекаем большую выгоду из дескрипторов, предоставляющих основу для рационального анализа.

Приведем несколько примеров, показывающих широту применения данного метода. На рис. 26 представлено исследование химического разложения CH_4 метана в условиях ударной компрессии (высокие температура $T = 3300\ K$ и давление $p = 40{,}53\ ГПа$).

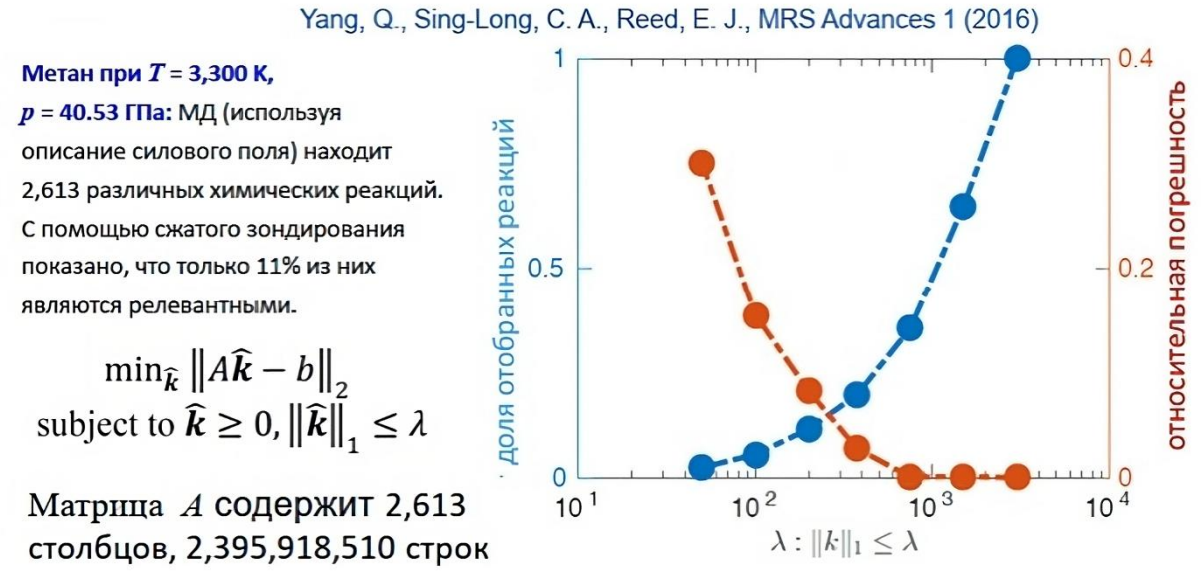


Рисунок 26. Химическое разложение CH_4 в условиях ударной компрессии

Для моделирования данного процесса использовалась молекулярная динамика, на основе которой было найдено 2613 различных химических процессов. Авторы задали вопрос: «А все ли реакции нужны? Насколько они важны?». Было использовано сжатое зондирование для определения релевантных реакций. Система, которую они минимизировали, включала популяцию b различных обломков метана - различных промежуточных соединений, по которым были найдены коэффициенты, минимизирующие разность энергий между предсказанной популяцией и той, что получили из молекулярной динамики. Установлено, что всего 11% реакций необходимо учесть для построения хорошей модели этого процесса.

Другой пример, который также интересен с точки зрения термоэлектриков, состоит в возможности использования сжатого зондирования для расчетов ангармоничности решетки и ее теплопроводности.

Авторы данной работы (рис. 27) рассчитали силы, действующие на атомы для различных смещений атомов. Сила представляется в виде разложения по степеням:

$$F_a = -\Phi_a - \Phi_{ab}u_b - \frac{1}{2}\Phi_{abc}u_b u_c - \cdots$$

Известны силы и смещения атомов. Из этих данных находятся коэффициенты ангармоничности *a*, *b*, *c*, то есть решается задача. Далее, с применением регуляризации LASSO, минимизируется ошибка по отношению к коэффициентам. Таким образом получается прогностическая модель динамики ангармонической решетки.

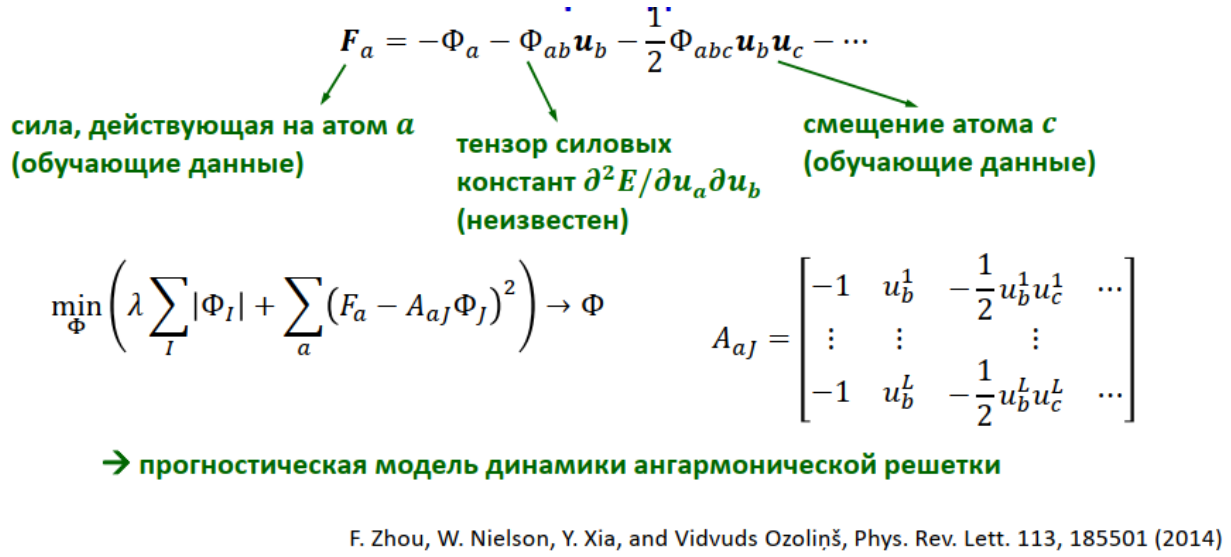
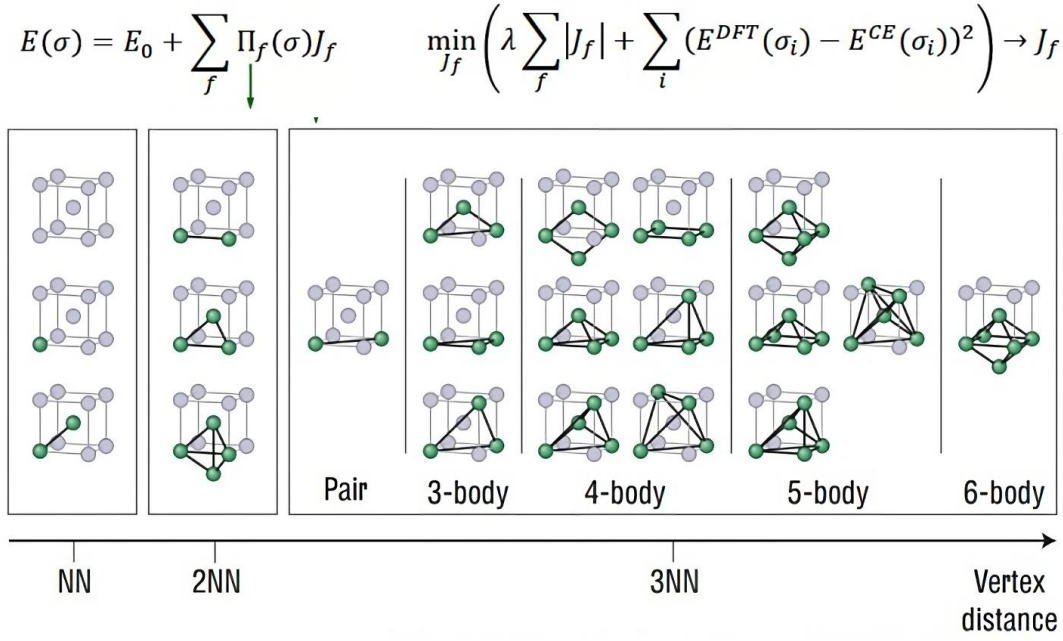


Рисунок 27. Определение ангармоничности решетки и ее теплопроводности с использованием метода сжатого зондирования

Также показателен пример применения метода сжатого зондирования для кластерного разложения (рис. 28).



L. J. Nelson, G. L. W. Hart, F. Zhou, and V. Ozoliņš, Phys. Rev. B 87, 035125 (2013)

Рисунок 28. Иллюстрация применения метода сжатого зондирования для кластерного разложения

Существуют системы, представленные в виде четко определенной решетки, в узлах которой можно менять тип атомов или создавать вакансию, которая будет рассматриваться как новый тип. Для предсказания энергии таких систем применяется метод разложение по кластерам, где энергия представляется в виде суммы энергий одиночных атомов и энергий многочастичных взаимодействий (для каждого взаимодействия находятся коэффициенты).

Авторы использовали LASSO для нахождения физически релевантных кластеров с наилучшей точностью. LASSO позволяет найти модель с наименьшим количеством ненулевых коэффициентов, что позволяет избежать переопределения и необходимости производить расчет с помощью дорогих методов для огромного числа тренировочных данных.

LASSO - простая задача оптимизации, эквивалентная полному перебору при условии, что признаки не коррелируют друг с другом. Чем больше производных признаков (математических комбинаций из основных признаков) мы сделаем, тем больше между ними будет корреляция. В таком случае LASSO перестает работать и применяется метод Sure Independence Screening + Selection Operator (SISSO) (рис. 29).

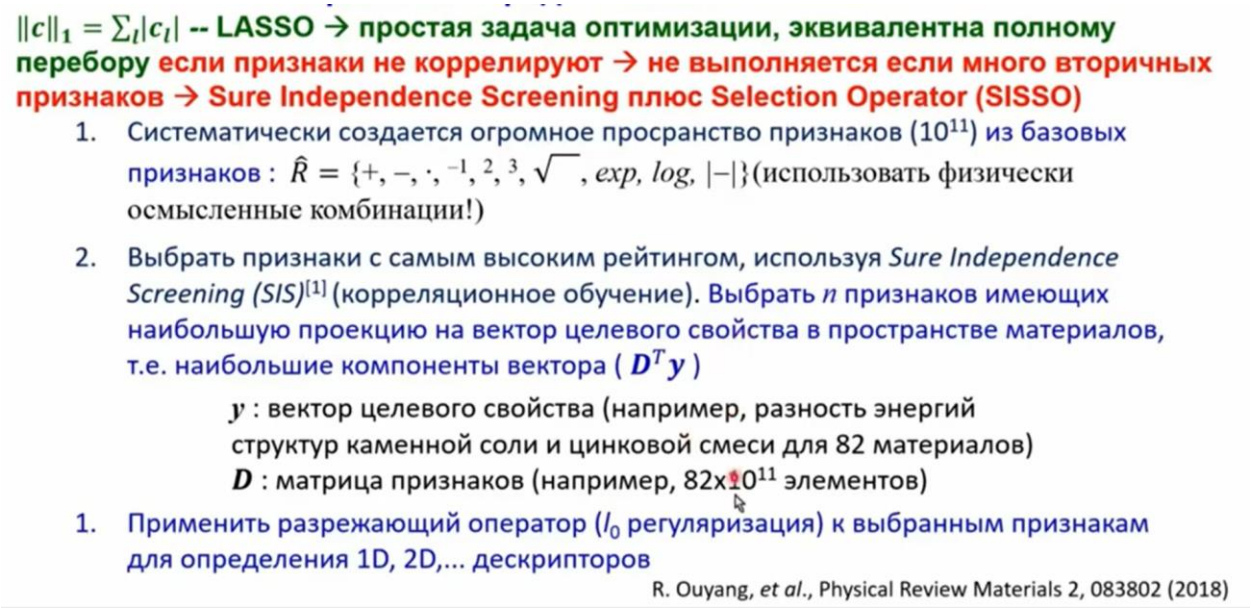


Рисунок 29. Метод Sure Independence Screening + Selection Operator (SISSO)

В SISSO создается огромное пространство признаков из базовых признаков с помощью математических операторов. Затем выбираются признаки с самым высоким рейтингом с помощью корреляционного обучения (Sure Independence Screening (SIS)) (смысл его прост: выбираются *n* признаков, имеющих наибольшую проекцию на вектор целевого свойства в пространстве материалов, то есть наибольшие компоненты вектора (D^Ty), гд y - вектор целевого свойства, D^T: транспонированная матрица признаков). Таким образом, на первом этапе мы определяем, какие дескрипторы лучше всего коррелируют с целевым свойством. На основании корреляции отбирается определенный набор и из него, с помощью l_0 регуляризации (полного перебора

данных, ранее отобранных с помощью корреляционного обучения), отбираются дескрипторы (рис. 30).

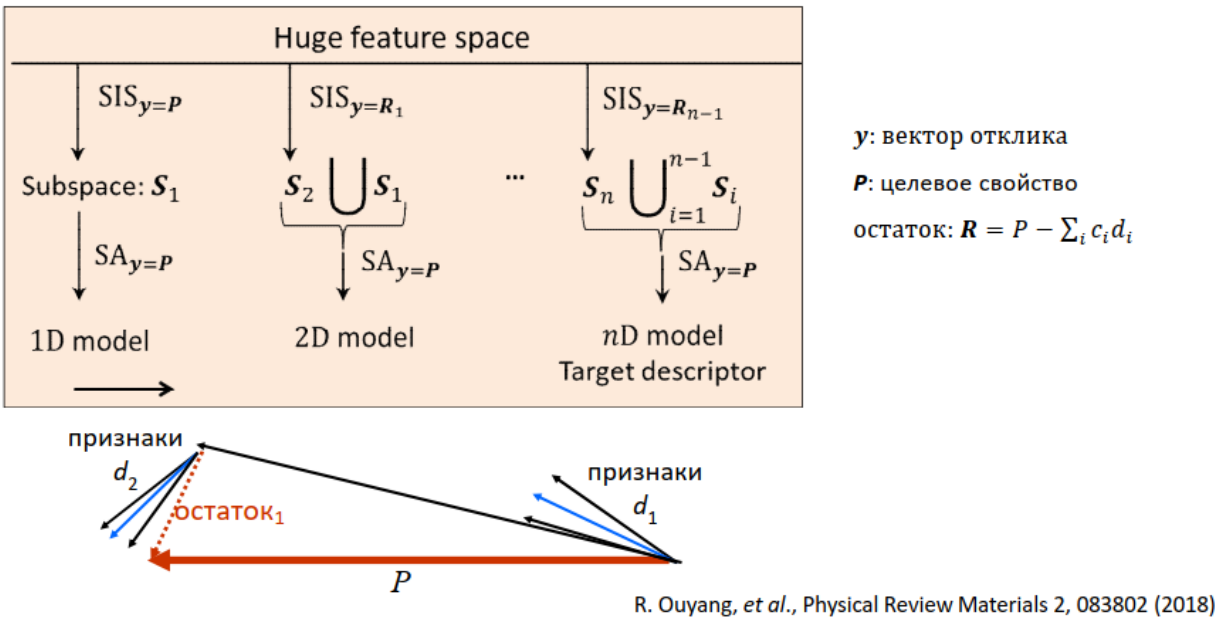


Рисунок 30. SISSO: Итеративная остаточная подгонка

Сперва отбираются дескрипторы, которые больше всего коррелируют с самим свойством, затем - с остатком, то есть с тем, что остается после того, как мы нашли модель и вычли ее из всего набора. Так делается шаг за шагом, и, таким образом, исследуется все пространство сложных признаков, полученных с помощью математических операторов.

На рис. 31 представлено сравнение эффективности методов LASSO и SISSO. Согласно рис. 31, наблюдается улучшение модели. Полученный дескриптор не сильно отличается по сложности от предыдущего, однако позволяет более четко распределять данные по областям.

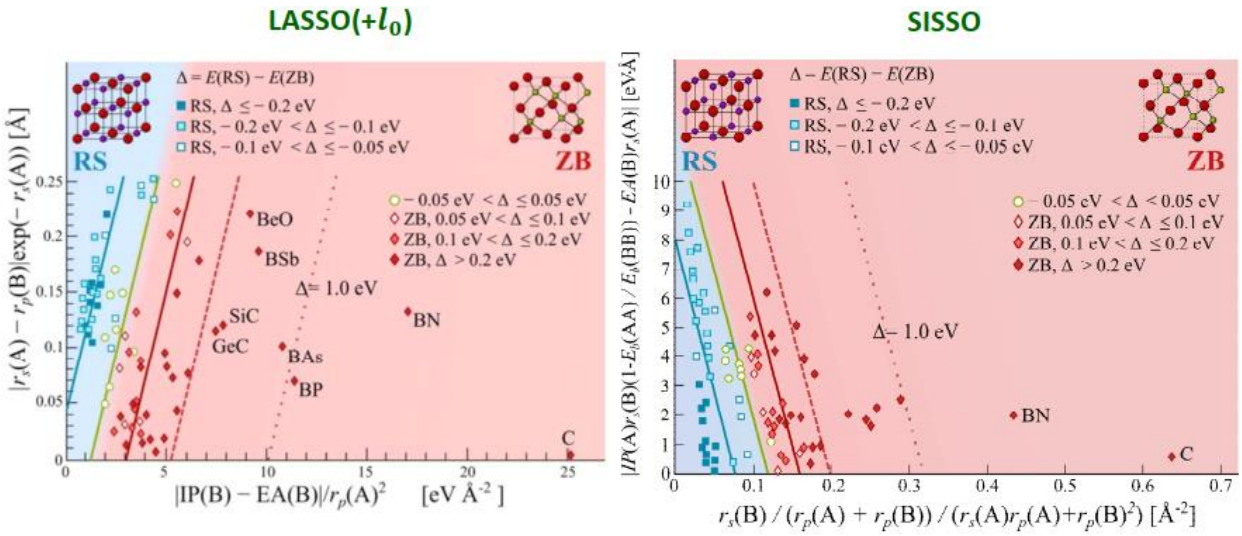


Рисунок 31. Сравнение эффективности методов LASSO и SISSO

Как видно из рис. 32, средние и максимальные ошибки в SISSO гораздо меньше (синяя и красная линия для двух разных уровней сложности), чем у других методов (LASSO, Eureqa).

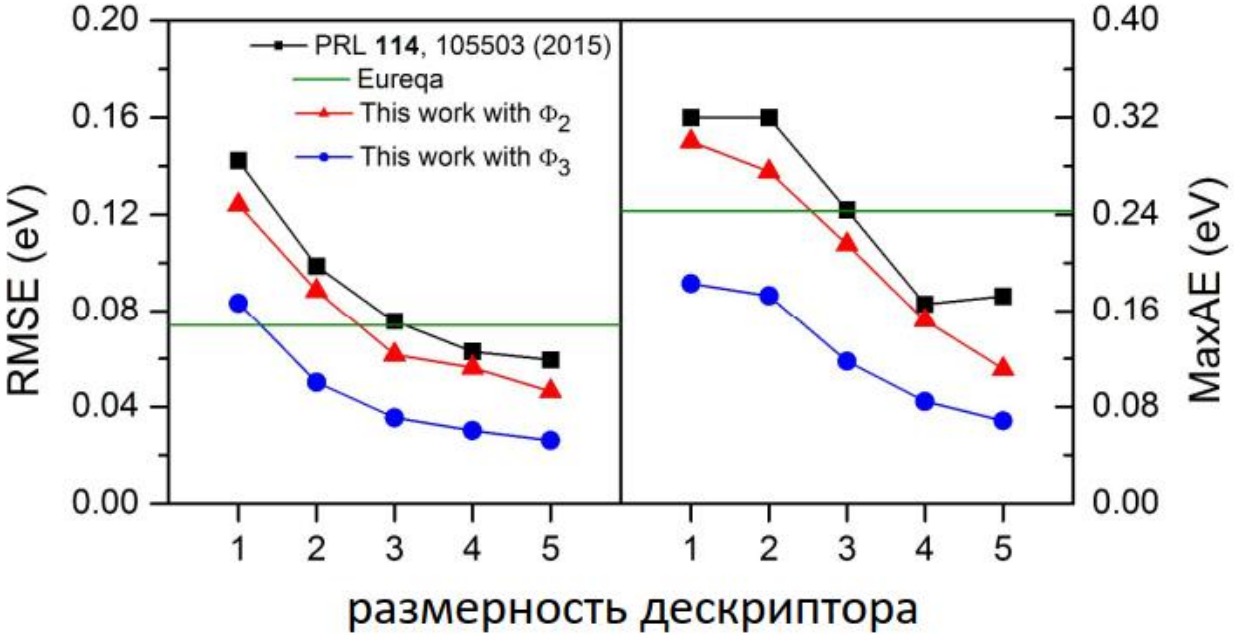


Рисунок 32. Сравнение эффективности различных методов

Метод SISSO можно применять также для мультизадачного поиска, когда есть несколько целевых свойств, которые мы хотим одновременно смоделировать с помощью одних и тех же дескрипторов (рис. 33).

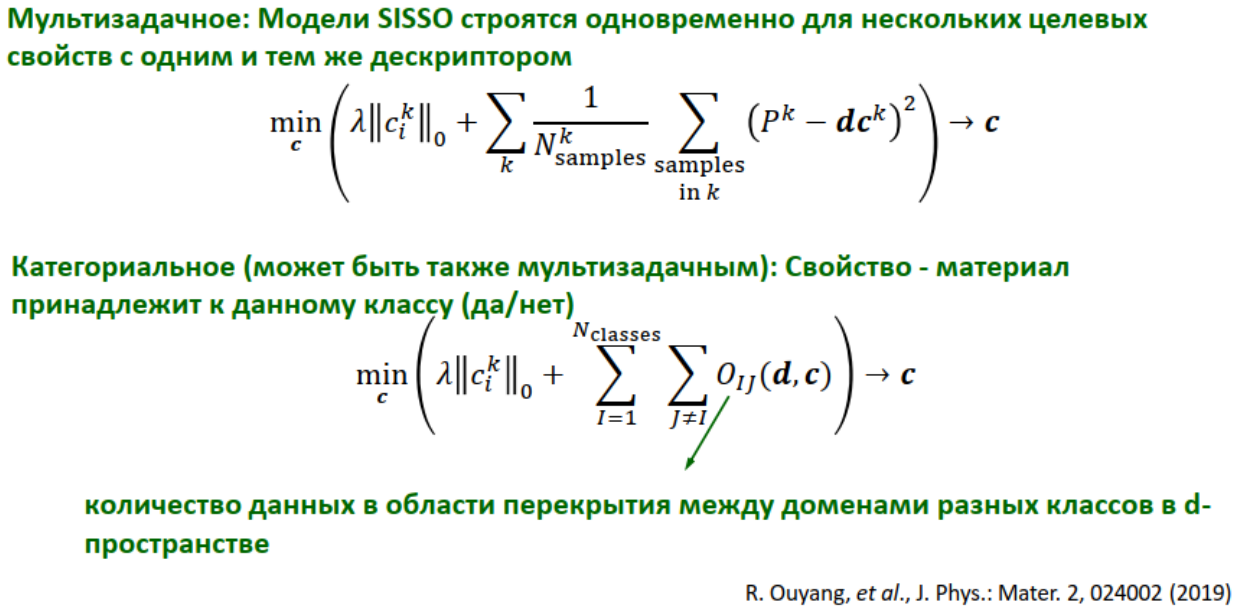


Рисунок 33. Метод SISSO: мультизадачный поиск

Также можно решать категориальные задачи, когда целевое свойство - определить, принадлежит ли материал к данному классу (ответ: да или нет), т.е. нет количественной оценки (в методе SISSO минимизируется перекрытие между областями, ограничивающими данный класс). Это применяется для предсказания кристаллических структур.

Интересным является пример предсказания стабильности фазы перовскита ABX_3 (рис. 34). Такая структура довольно стабильна. Множество комбинаций атомов A и B сохраняют эту структуру, что удобно для рационального дизайна материалов (структура сохраняется, значит свойства зависят от состава).

Можно ли предсказать, какие пары или даже тройки атомов образуют такую структуру? Гольдшмидт предложил критерий стабильности структуры перовскитной фазы, однако он не всегда дает правильную оценку. Значит, его необходимо улучшить.

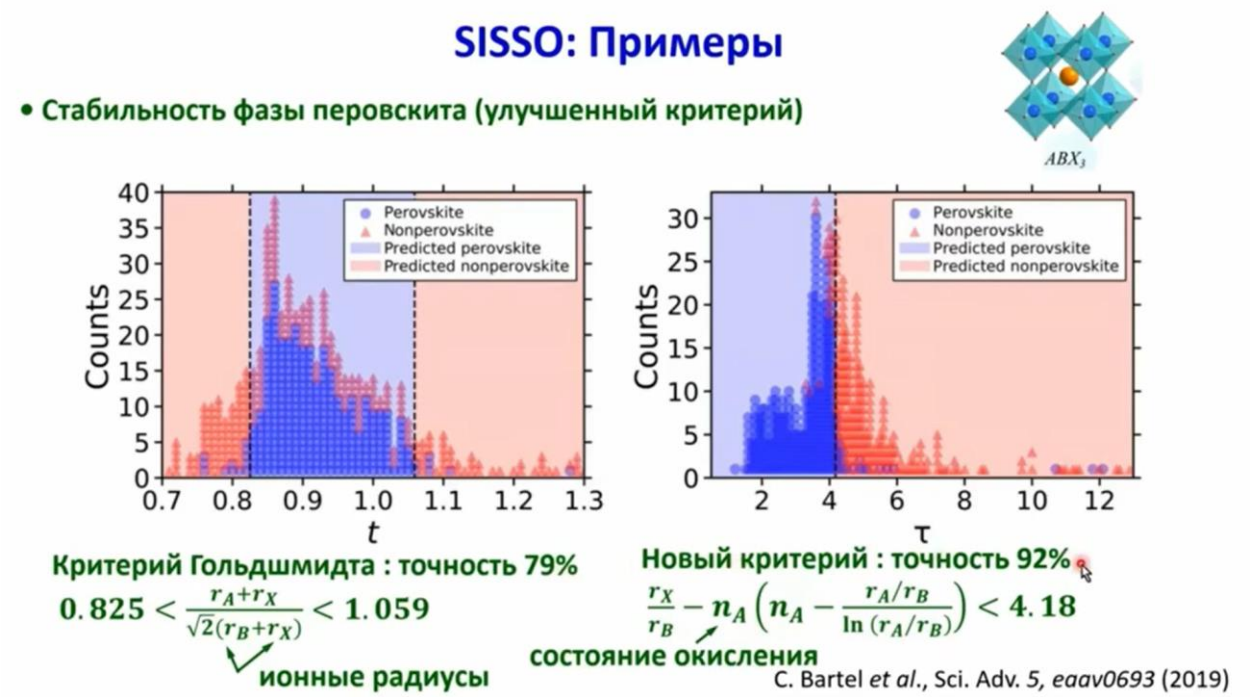


Рисунок 34. Пример предсказания стабильных фаз

Поэтому C. Bartel в своей работе с помощью символической регрессии SISSO определил другой дескриптор, учитывающий не только радиусы, но и состояния окисления, что сильно повысило точность: более 90% по сравнению с критерием Гольдшмидта (79%).

В следующей работе использовался метод многоцелевого SISSO, когда несколько свойств определяются одним дескриптором (рис. 35). Изучалась адсорбция молекул (C, CH, CO, H, O, OH) на металлических поверхностях (не только чистых металлов, но и сплавов разных типов), что важно для катализа. На рис. 35 представлены ошибки различных моделей предсказания.

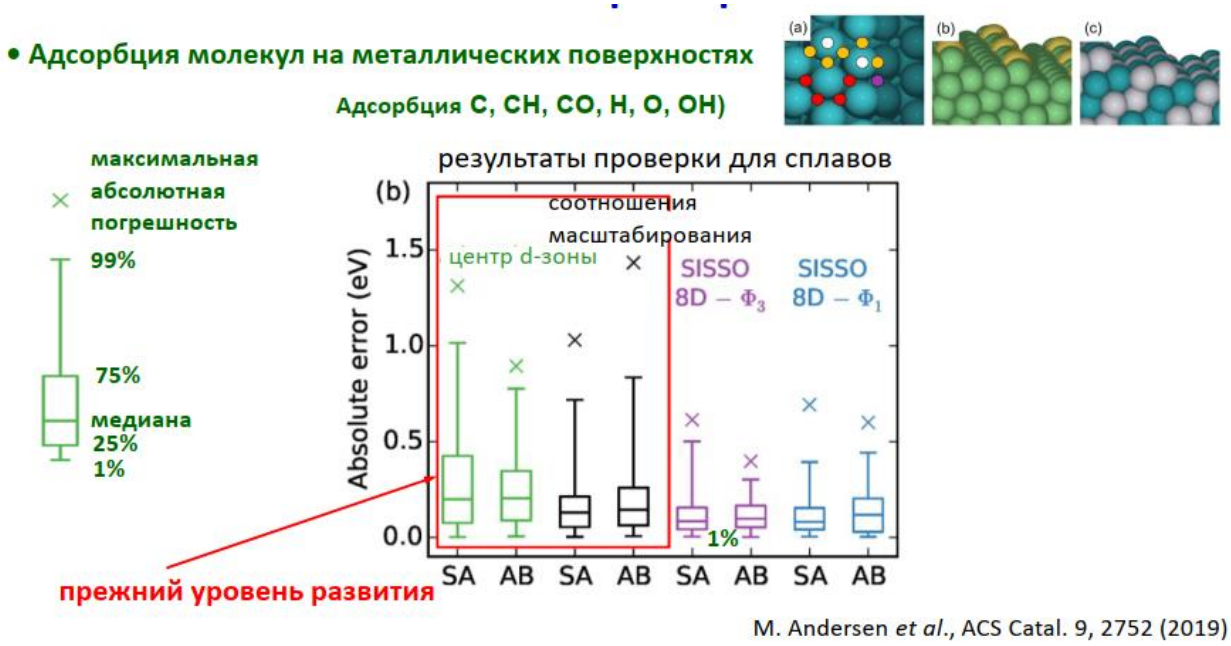


Рисунок 35. Ошибки различных моделей предсказания

Наглядно показано, что анализ данных искусственным интеллектом с помощью сжатого зондирования методом SISSO дает хорошую точность.