Научно-технологический университет Сириус

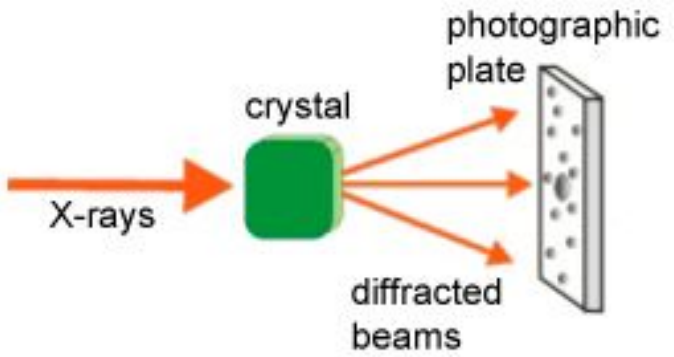
Структурная биоинформатика | Лекция 5

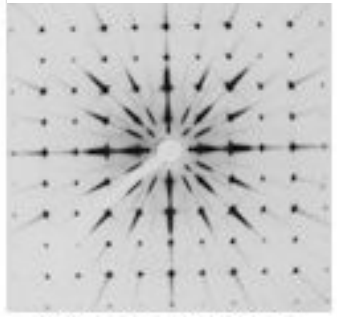
Оценка качества построения модели

Александр Злобин

РСА эксперимент

1. Получение кристалла белка
2. Получение дифракционной картины
3. Вычисление модулей структурных факторов (F)
4. Определение фаз (решение фазовой проблемы) (φ)
5. Построение функции электронной плотности в ячейке
6. Интерпретация электронной плотности, построение черновой модели
7. Сравнение электронной плотности модели с экспериментальной, оптимизация модели
8. Финальная модель



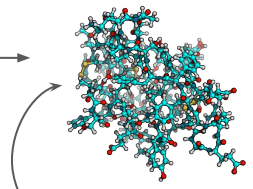


→ **F**

SIR/MIR/AD/MAD/... → **φ**

F, φ → **ρ(x,y,z)**

ρ(x,y,z) →



opt

Где можно ошибиться?

1. Получение кристалла белка
2. Получение дифракционной картины
3. Вычисление модулей структурных факторов (F)
4. Определение фаз структурных факторов (φ)
5. Построение функции электронной плотности в ячейке
6. Интерпретация электронной плотности, построение черновой модели
7. Сравнение электронной плотности модели с экспериментальной, оптимизация модели
8. Финальная модель

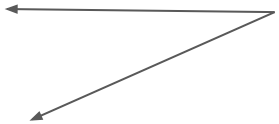
Nunn et al., 1995.
Бактериоферритин вместо LH1

Hoier et al.,1994.
Ошибка в определении группы симметрии

‘great pentaretraction’
5 статей по ABC-транспортерам.
I(h, k, l) -> |F(−h, −k, −l)|,
I(−h, −k, −l) -> |F(h, k, l)|

Где можно ошибиться?

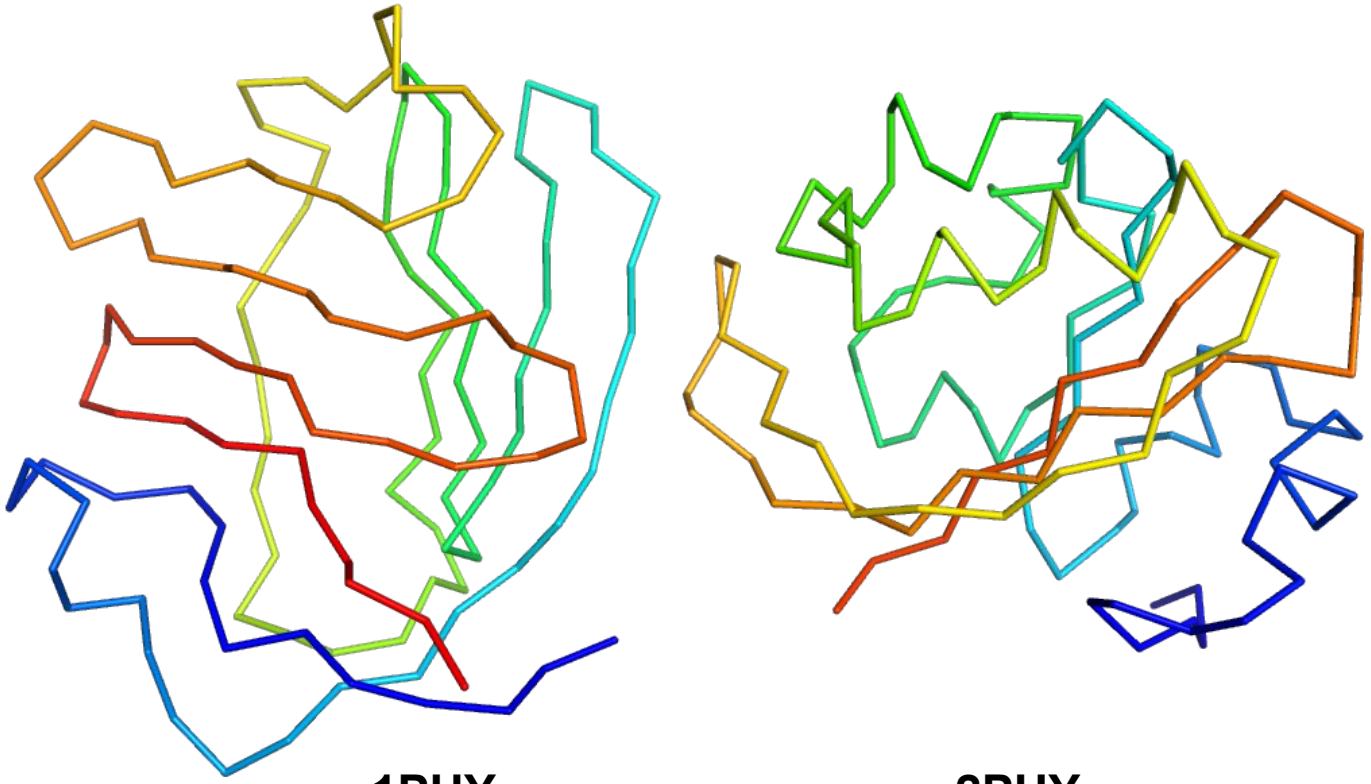
1. Получение кристалла белка
2. Получение дифракционной картины
3. Вычисление модулей структурных факторов (F)
4. Определение фаз структурных факторов (φ)
5. Построение функции электронной плотности в ячейке
6. Интерпретация электронной плотности, построение черновой модели
7. Сравнение электронной плотности модели с экспериментальной, оптимизация модели
8. Финальная модель



Подвержены субъективному влиянию исследователя

Примеры ошибок

Photoactive yellow protein: структура полностью ошибочна

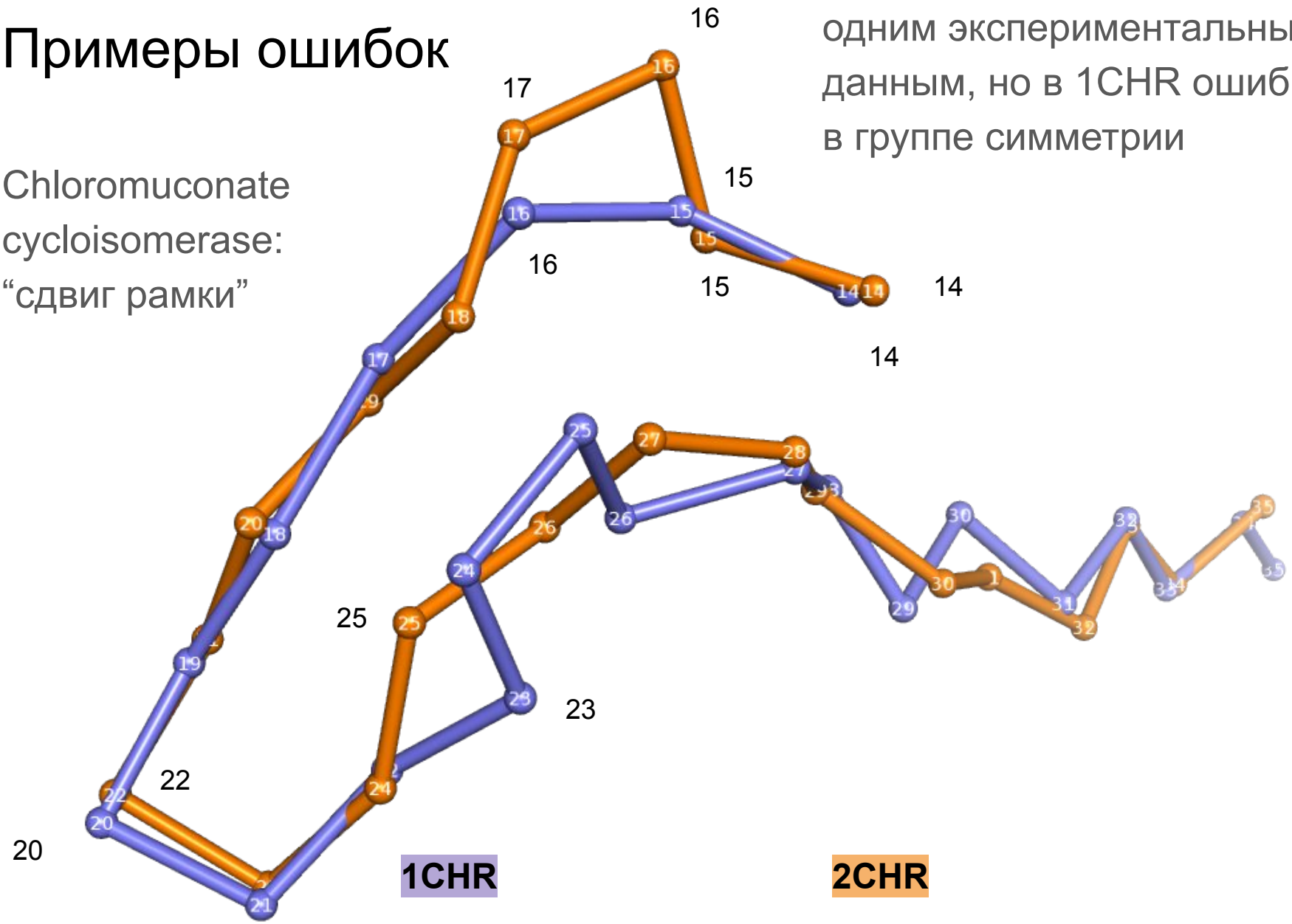


1PHY **2PHY**

Примеры ошибок

Chloromuconate cycloisomerase: “сдвиг рамки”

Модели построены по одним экспериментальным данным, но в 1CHR ошибка в группе симметрии



Примеры ошибок

Модели построены по одним экспериментальным данным, но в 1CHR ошибка в группе симметрии

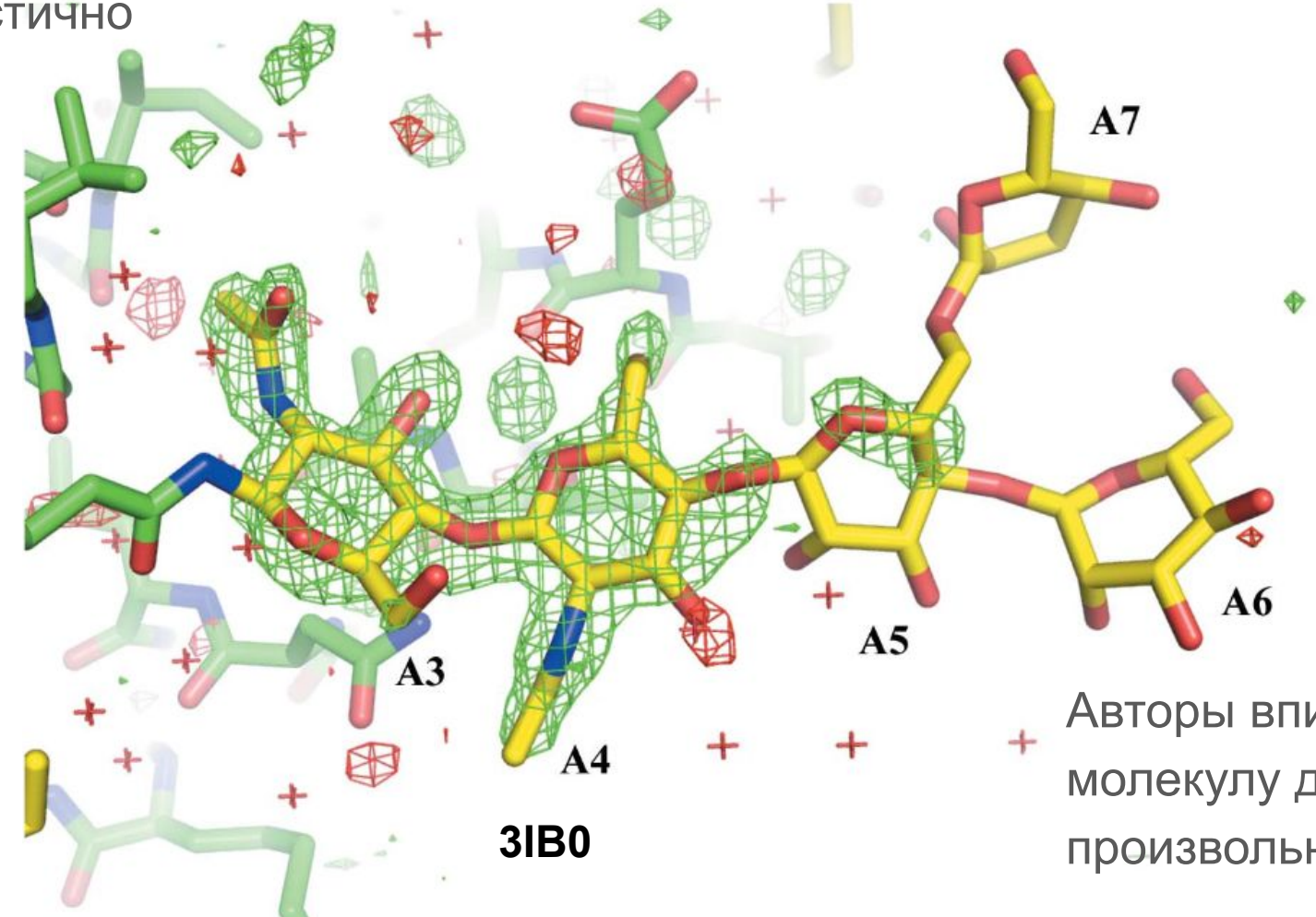
Chloromuconate cycloisomerase:
“сдвиг рамки”

```
                 *         20          *         40          *         60          *
1chr : MKIDAIEAVIVDVPT-KR-PIQMSITTVHQQSYVIVRVYSEGLVGVGEGGSVGGPVWSAECAETIKIIVERYL : 71
2chr : MKIDAIEAVIVDVPTKRPIQMSITT-VHQQSYVIVRVYSE-GLVGVGEGGSVGGPVWSAECAETIKIIVERYL : 71

          80          *         100
1chr : APHLLGTDAFNVSGALQTMARAVTGNA : 98
2chr : APHLLGTDAFNVSGALQTMARAVTGNA : 98
```

Примеры ошибок

Гликозилированные остатки лактотрансферрина: модель вписана в экспериментальную электронную плотность лишь частично

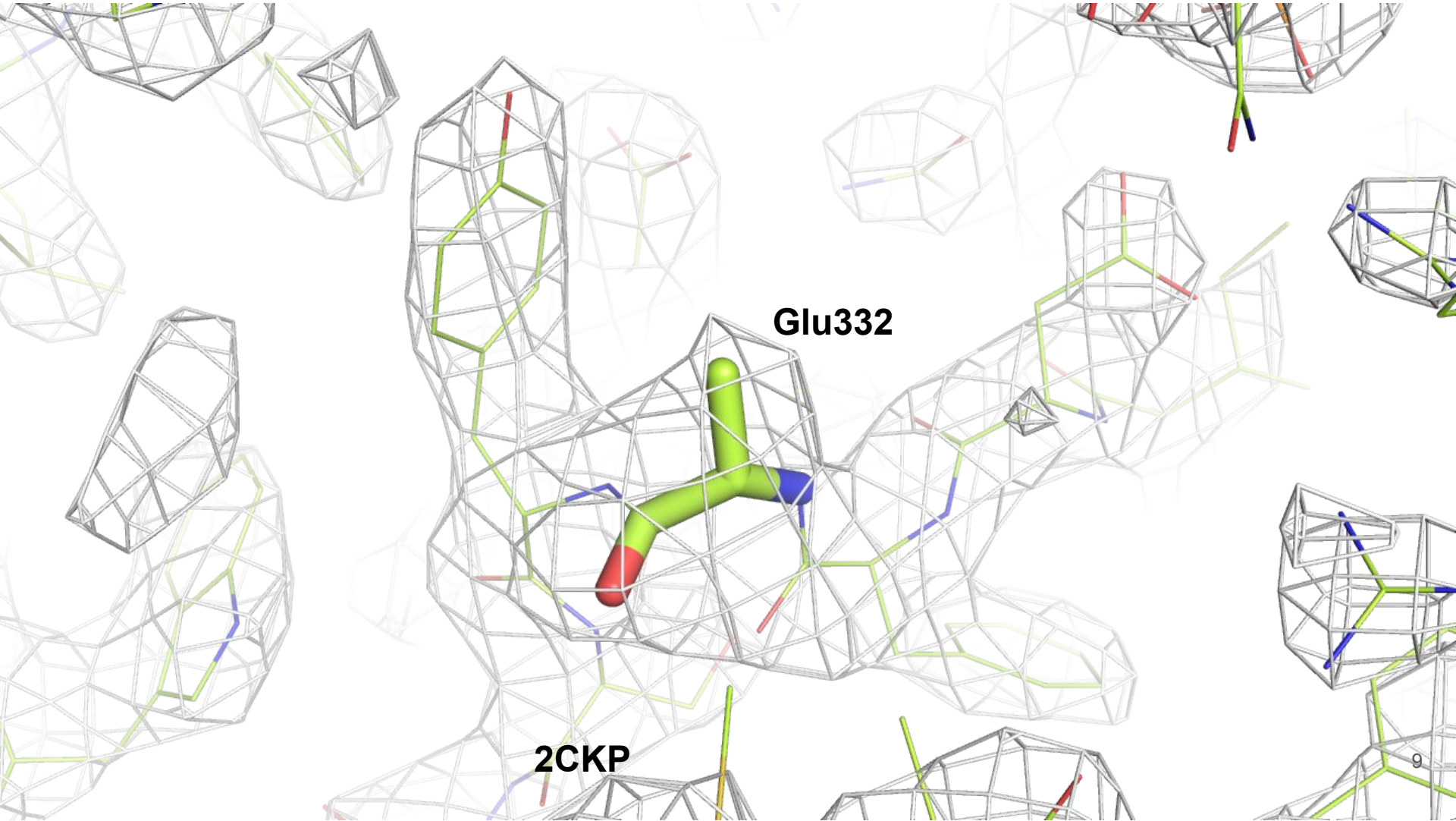


3IB0

Авторы вписывают молекулу достаточно произвольно

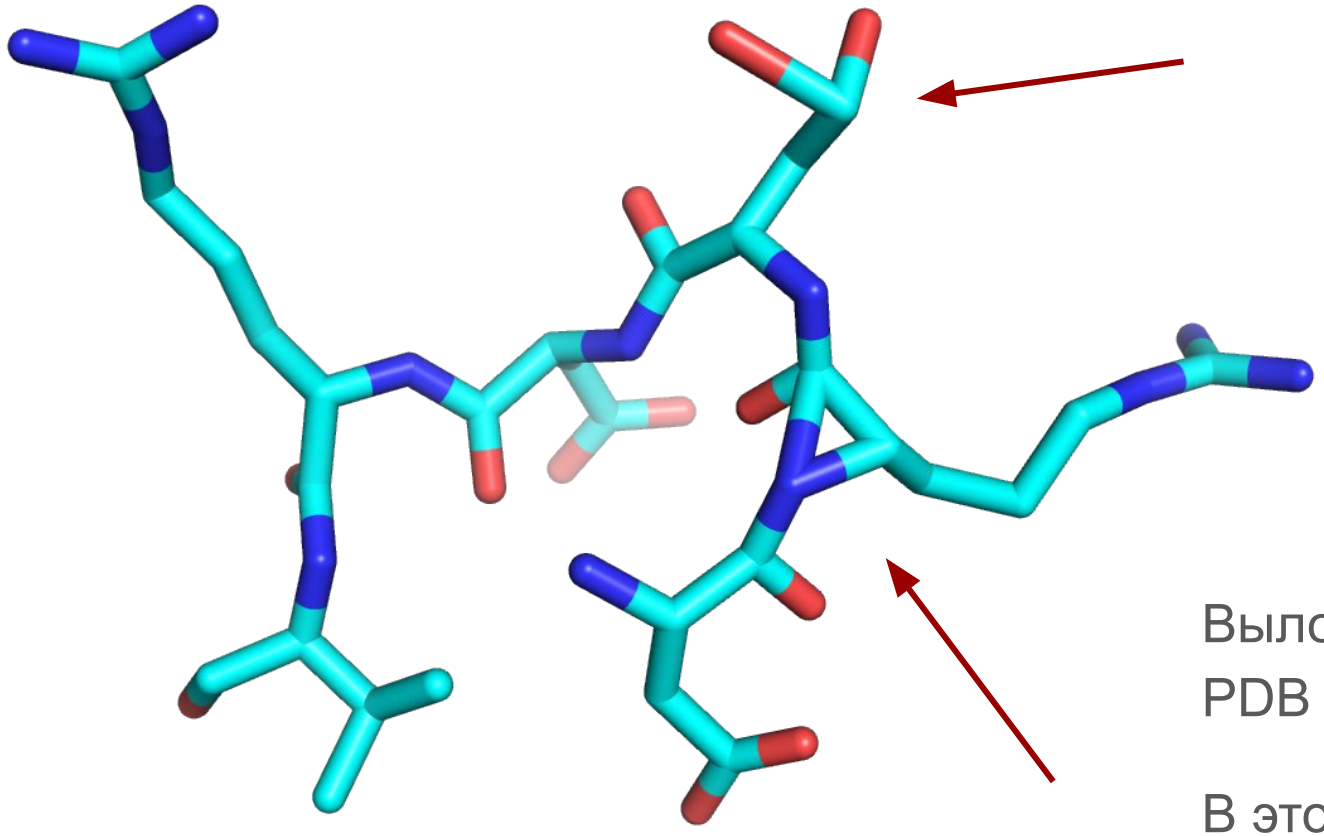
Примеры ошибок

Недоразрешённые остатки



Примеры ошибок

Fetuin-binding protein: модель имеет неадекватные с точки зрения химии длины связей и углы



1DLP

Выложена в PDB в 1999г.

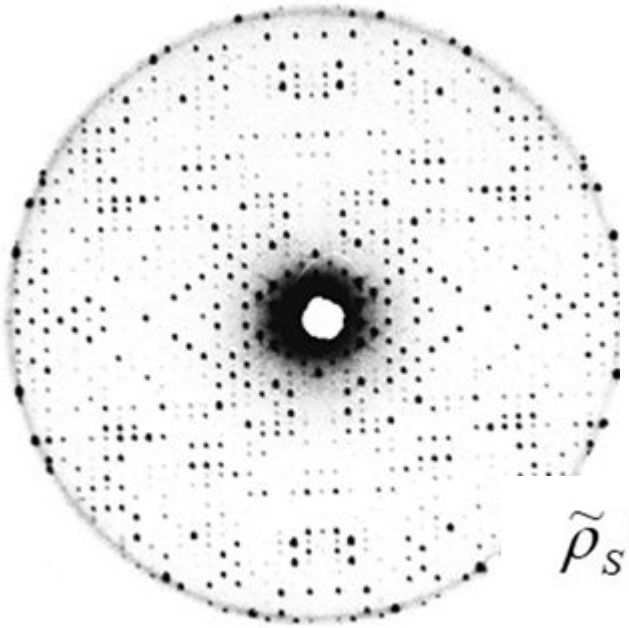
В этом году юбилей)

Validation - контроль качества полученных моделей

- Контроль качества экспериментальных данных
- Контроль соответствия модели и экспериментальной электронной плотности
- Контроль соответствия модели нашим представлениям о физико-химических свойствах молекул

Показатели качества модели в целом: Разрешение

- Вычисляется по набору структурных факторов и параметрам ячейки
- Характеризует совокупность экспериментальных данных – структурных факторов



Решение фазовой проблемы

I(h,k,l) → **F(h,k,l)** ← **φ(h,k,l)**

$$\tilde{\rho}_S(x,y,z) \approx \sum_{(hkl)\in S} F_{hkl} \cos\left[2\pi(hx+ky+lz)-\varphi_{hkl}\right]$$

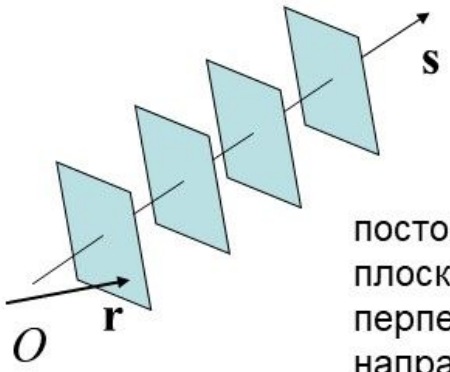
Показатели качества модели в целом: Разрешение

Гармоника Фурье

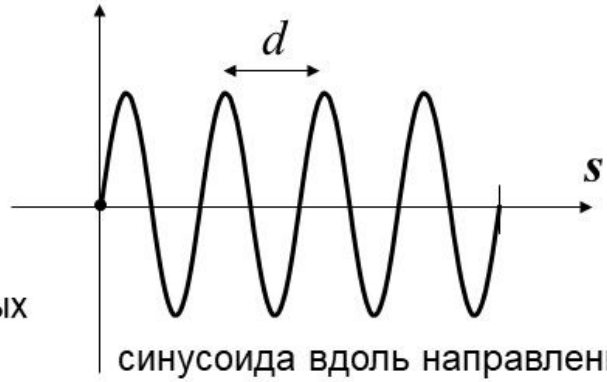
$$H_{\mathbf{s}}(\mathbf{r}) = \cos\left[2\pi(hx + ky + lz)\right] = \cos\left[2\pi(\mathbf{s}, \mathbf{r})\right]$$

$$\mathbf{r} = x\mathbf{a} + y\mathbf{b} + z\mathbf{c}$$

$$h = (\mathbf{s}, \mathbf{a}), k = (\mathbf{s}, \mathbf{b}), l = (\mathbf{s}, \mathbf{c})$$



постоянна на плоскостях, перпендикулярных направлению **s**

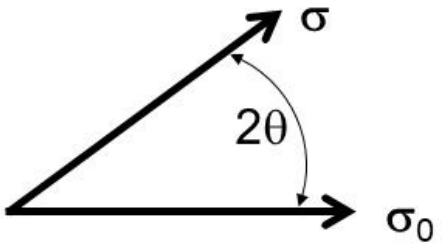


синусоида вдоль направления **s**

Разрешение, соответствующее гармонике Фурье

$$d = \frac{1}{|\mathbf{s}|}$$ - расстояние между соседними максимумами в направлении **s**;

$$|\mathbf{s}| = \frac{2\sin\theta}{\lambda} \qquad d = \frac{\lambda}{2\sin\theta}$$



Показатели качества модели в целом: Разрешение

- Разрешение полного набора гармоник равно минимальному разрешению из всех гармоник набора
- В эксперименте получается измерить не все гармоники
- Полнота данных для данного разрешения - % гармоник с большим разрешением, которые удалось измерить в эксперименте

Измеренный рефлекс

Потерянный рефлекс

Разрешение для данного набора можно принять таким:
1 Å с полнотой данных 20/32*100% = 62%
1.5 Å с полнотой 19/22 *100% = 86%
2 Å с полнотой 18/20 *100% = 90%

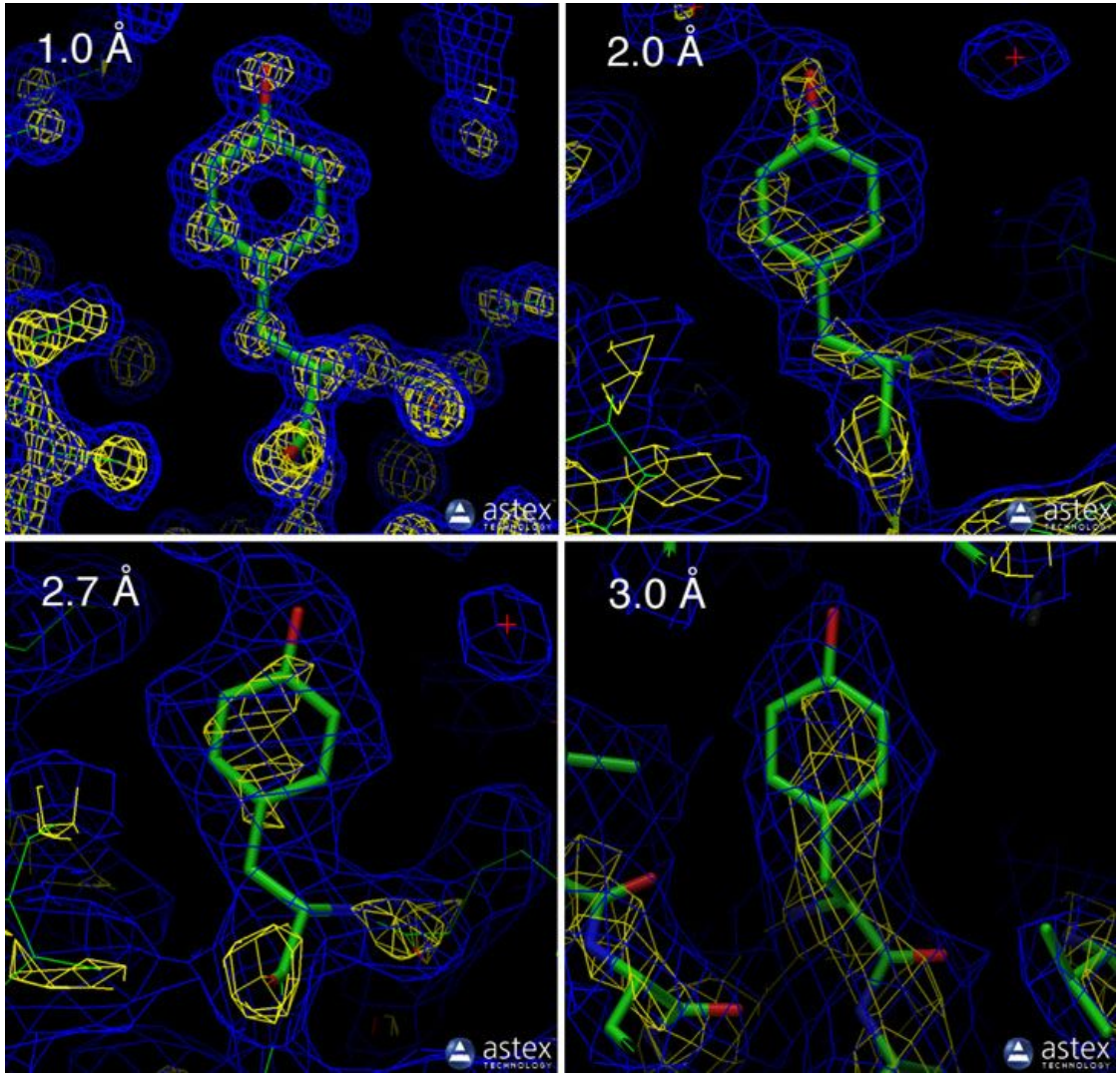
Показатели качества модели в целом: Разрешение

- **Разрешение структуры** определяется **разрешением измеренной гармоники** с самым малым разрешением и **полнотой** измеренного набора
- Есть небольшая доля субъективности в определении разрешения

```
REMARK   3    REFINEMENT TARGET : MAXIMUM LIKELIHOOD
REMARK   3
REMARK   3  DATA USED IN REFINEMENT.
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) : 1.36
REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) : 38.32
REMARK   3   DATA CUTOFF             (SIGMA(F)) : NONE
REMARK   3   COMPLETENESS FOR RANGE        (%) : 98.52
REMARK   3   NUMBER OF REFLECTIONS             : 89715
REMARK   3
```

Показатели качества модели в целом: Разрешение

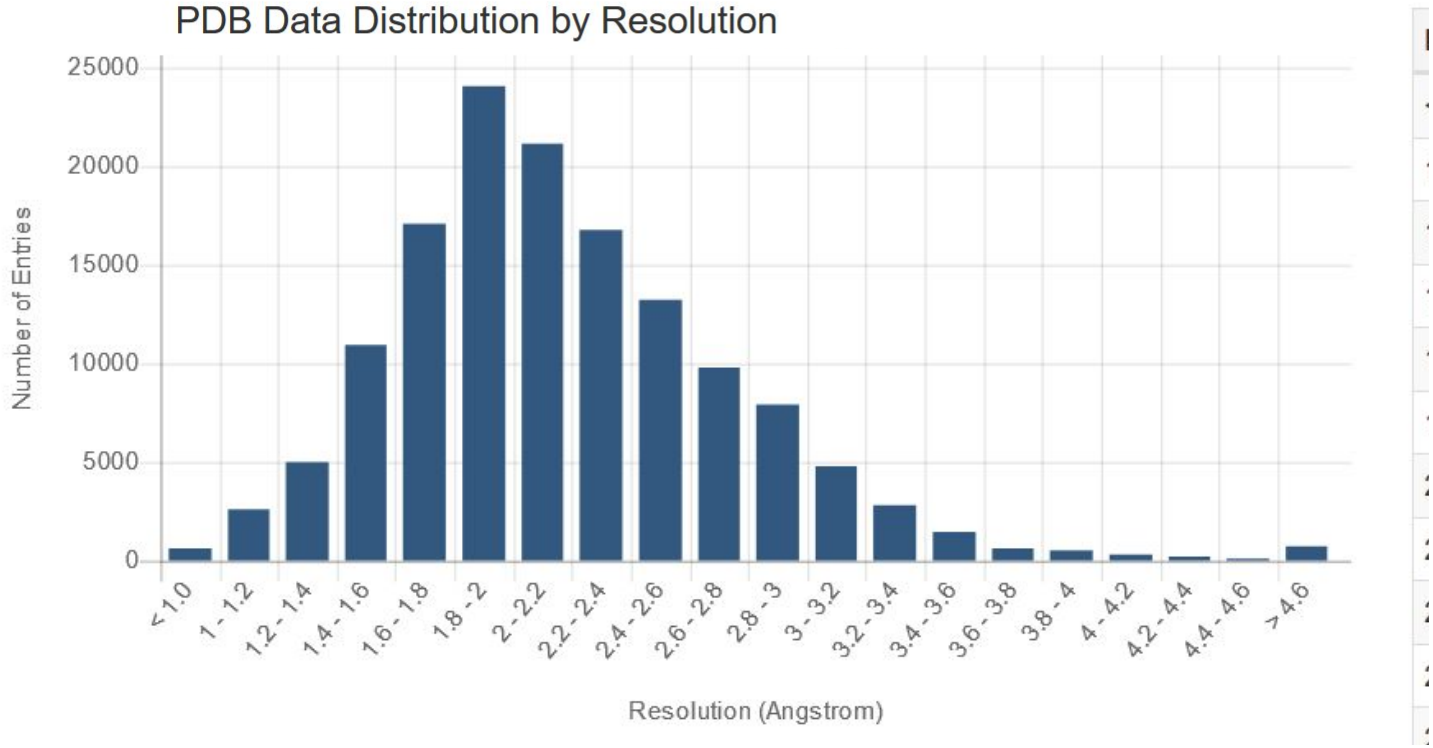
- Чем лучше - меньше - разрешение (в Å), тем менее вероятны ошибки
- По данным РСА с плохим разрешением можно построить хорошую модель; при хорошем разрешении в модели могут быть ошибки



Tyr103 миоглобина из структур с разным разрешением

Показатели качества модели в целом: Разрешение

Высокое разрешение: <1.5 Å
Хорошее разрешение: 1.5 – 2.5 Å
Удовлетворительное: 2.5 – 3.5 Å
Низкое: > 3.5 Å

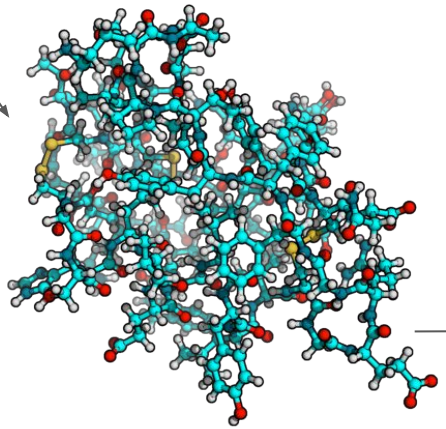


Показатели качества модели в целом: R-фактор

- Вычисляется по оптимизированной модели и измеренным структурным факторам
- Характеризуют соответствие модели экспериментальным данным – структурным факторам

Fobs(h,k,l) - измерены в эксперименте

$$R = \frac{\sum_{hkl} \left| F_{hkl}^{calc} - F_{hkl}^{obs} \right|}{\sum_{hkl} F_{hkl}^{obs}} * 100\%$$



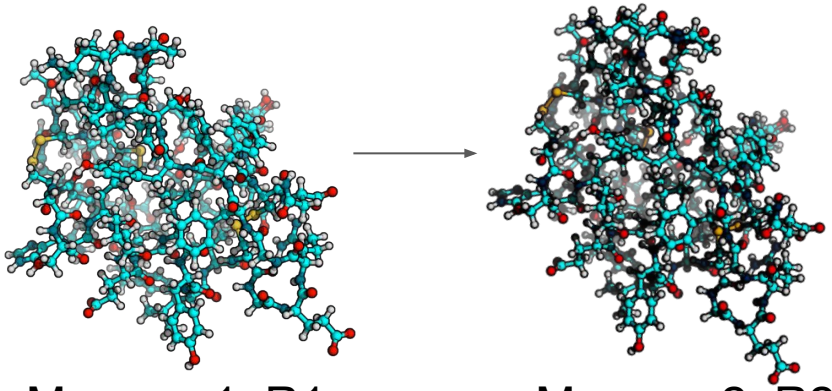
модель

Fcalc(h,k,l) - рассчитаны из модели

Показатели качества модели в целом: R-фактор

- В процессе оптимизации модели уточняется:
  - соответствие Fobs(h,k,l) и Fcalc(h,k,l)
  - соответствие длин связей и валентных углов модели физико-химическим свойствам молекул
- Минимизируем составной R-фактор:

$$R_{mixed} = w_{X\text{-}ray}R_{X\text{-}ray} + w_{dist}R_{dist} + w_{angle}R_{angle}$$



Модель 1, R1 Модель 2, R2

if R2 < R1 -> M2, else: M1

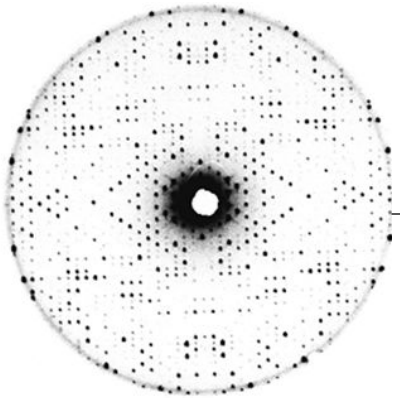
Показатели качества модели в целом: R-фактор

- Чем ниже R-фактор, тем больше структурные факторы модели соответствуют экспериментальным
- Проблема: перееоптимизация. Можно подогнать даже очень неправильную модель.

Table 1. Comparison of some model and refinement statistics of CRABP II (intentionally traced backwards; model X), CRABP I refined conservatively (with NCS constraints and grouped temperature factors; model Y), and the CRABP I structure refined according to today's 'standard' refinement practices (model Z).

Model	X	Y	Z
Resolution range (Å)	8.0–3.0	8.0–2.9	6.0–2.9
R	0.214	0.251	0.169

Показатели качества модели в целом: R_{free}



h	k	l	F
3	5	5	207.9
3	5	6	255.9
3	5	7	328.1
3	5	8	298.8
3	5	9	583.4
3	5	10	833.9
3	5	11	1403.0

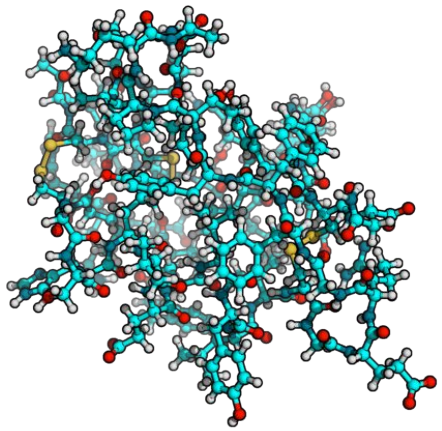
Набор измеренных рефлексов

h	k	l	F
3	5	5	207.9
3	5	6	255.9

Контрольный набор

h	k	l	F
3	5	7	328.1
3	5	8	298.8
3	5	9	583.4
3	5	10	833.9
3	5	11	1403.0

Рабочий набор



модель

Показатели качества модели в целом: R_{free}

- Вычисляется по той же формуле, что и R-фактор, но только для Fobs контрольных рефлексов и Fcalc финальной модели после оптимизации
- Если модель правильная, то R-free окажется примерно равным R-фактору или немногим больше
- Если R-free сильно больше R-фактора, модель переоптимизирована

$$R = \frac{\sum_{hkl} \left| F_{hkl}^{calc} - F_{hkl}^{obs} \right|}{\sum_{hkl} F_{hkl}^{obs}} * 100\%$$

- Хорошие значения: $R_{free} < 20\%$
- Плохие значения: $R_{free} > 40\%$
- Значения (R_{free} – R) > 10% настораживают в отношении переоптимизации (ovefitting)

Показатели качества модели в целом: R_{free}

```
REMARK   3
REMARK   3  FIT TO DATA USED IN REFINEMENT.
REMARK   3   CROSS-VALIDATION METHOD                 : THROUGHOUT
REMARK   3   FREE R VALUE TEST SET SELECTION         : RANDOM
REMARK   3   R VALUE     (WORKING + TEST SET)       : 0.15621
REMARK   3   R VALUE              (WORKING SET)     : 0.15185
REMARK   3   FREE R VALUE                           : 0.19471
REMARK   3   FREE R VALUE TEST SET SIZE   (%)       : 10.1
REMARK   3   FREE R VALUE TEST SET COUNT            : 5989
REMARK   3
```

- Авторы получили в эксперименте 59 297 рефлексов
- Они утверждают, что 10.1% данных спрятали в сейф
- По оставшимся 53 308 рефлексам оптимизировали модель и получили R=15%
- После этого достали тайные рефлексы из сейфа и рассчитали R-фактор по ним. Авторы получили R_{free} = 19%.

Показатели качества модели в целом: R_{free}

Table 1. Comparison of some model and refinement statistics of CRABP II (intentionally traced backwards; model X), CRABP I refined conservatively (with NCS constraints and grouped temperature factors; model Y), and the CRABP I structure refined according to today's 'standard' refinement practices (model Z).

Model	X	Y	Z
Resolution range (Å)	8.0–3.0	8.0–2.9	6.0–2.9
R	0.214	0.251	0.169
R_{free}	0.617	0.320	0.323

Показатели качества отдельных остатков

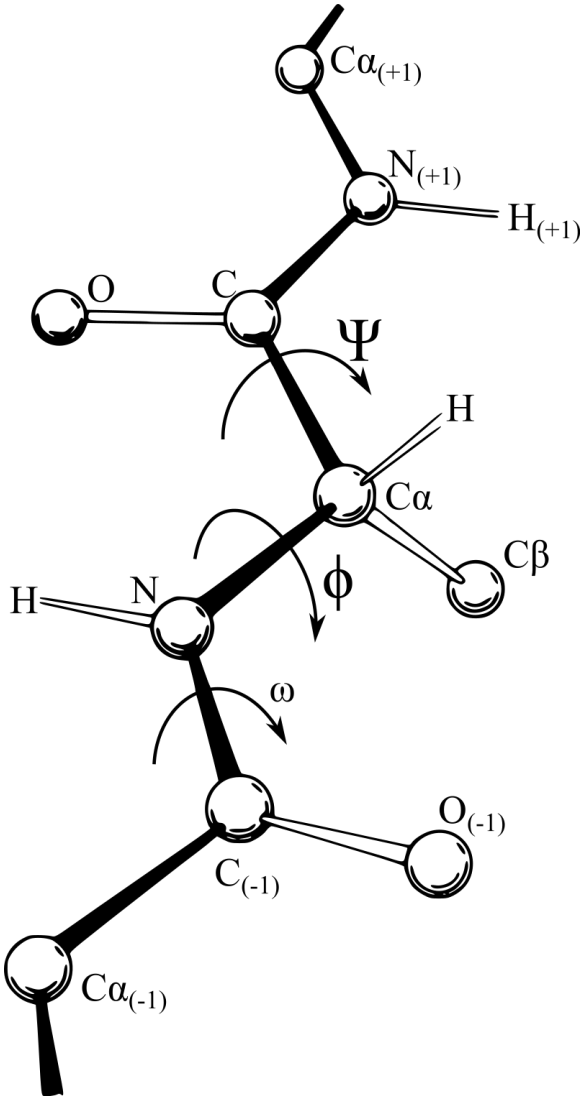
- Торсионные углы (Карты Рамачандрана и ротамеры боковых цепей)
- RSR
- Комфортность окружения остатка

Остатки с нетипичными значениями показателей - **маргинальные.**

Маргинальность остатка - ошибка расшифровки либо функционально значимая особенность.

% маргинальных остатков - ещё одна характеристика качества модели в целом

Показатели качества отдельных остатков: укладка полипептидной цепи



Некоторые значения геометрических параметров энергетически выгодны, и потому встречаются часто. При оптимизации модели геометрические параметры подгоняются под такие табличные значения.

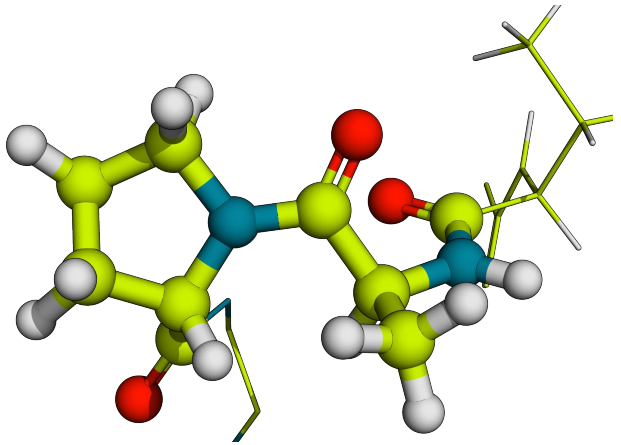
Укладка полипептидной цепи определяется тремя торсионными углами φ, ψ, ω

(Торсионный угол принимает значения от -180° до 180°)

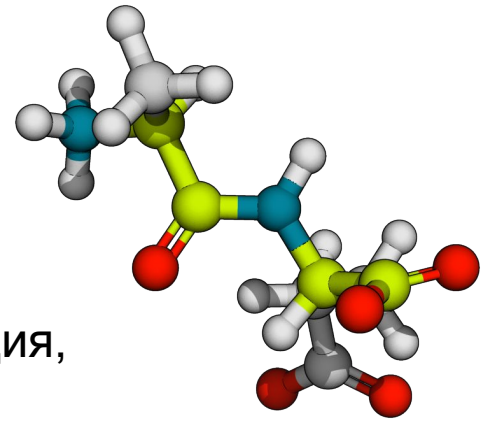
Показатели качества отдельных остатков: укладка полипептидной цепи

ω принимает всего два значения - 0° или 180° (почему?)

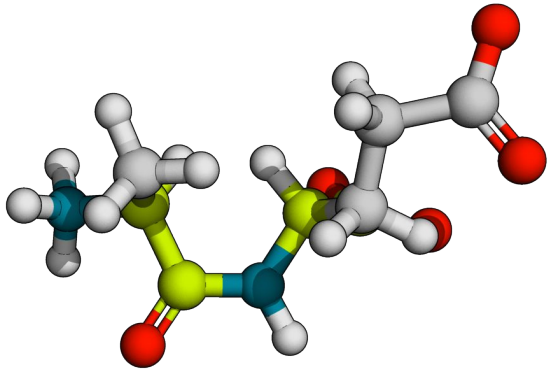
Транс-конформация наиболее частая для всех остатков кроме пролина.



Цис-конформация, ω=0°

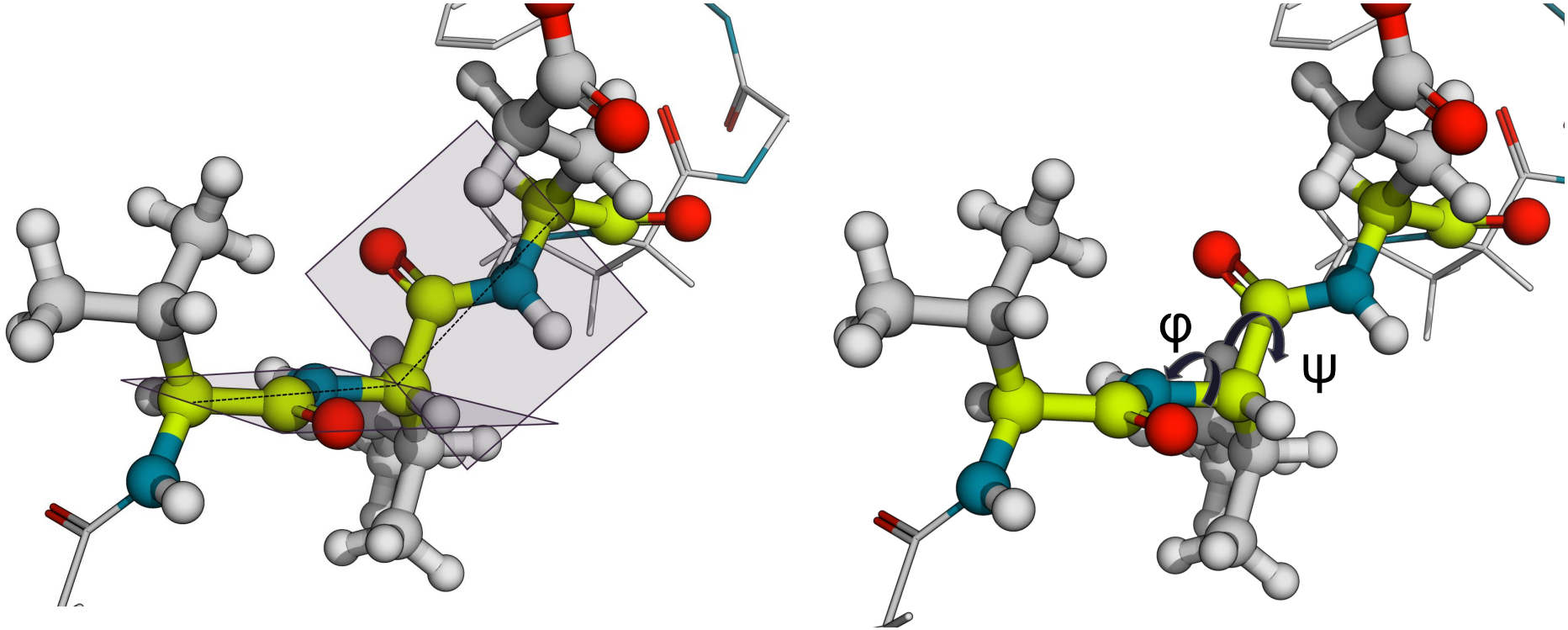


Транс-конформация, ω=180°



Цис-конформация, ω=0°

Показатели качества отдельных остатков: карты Рамачандрана для φ и ψ



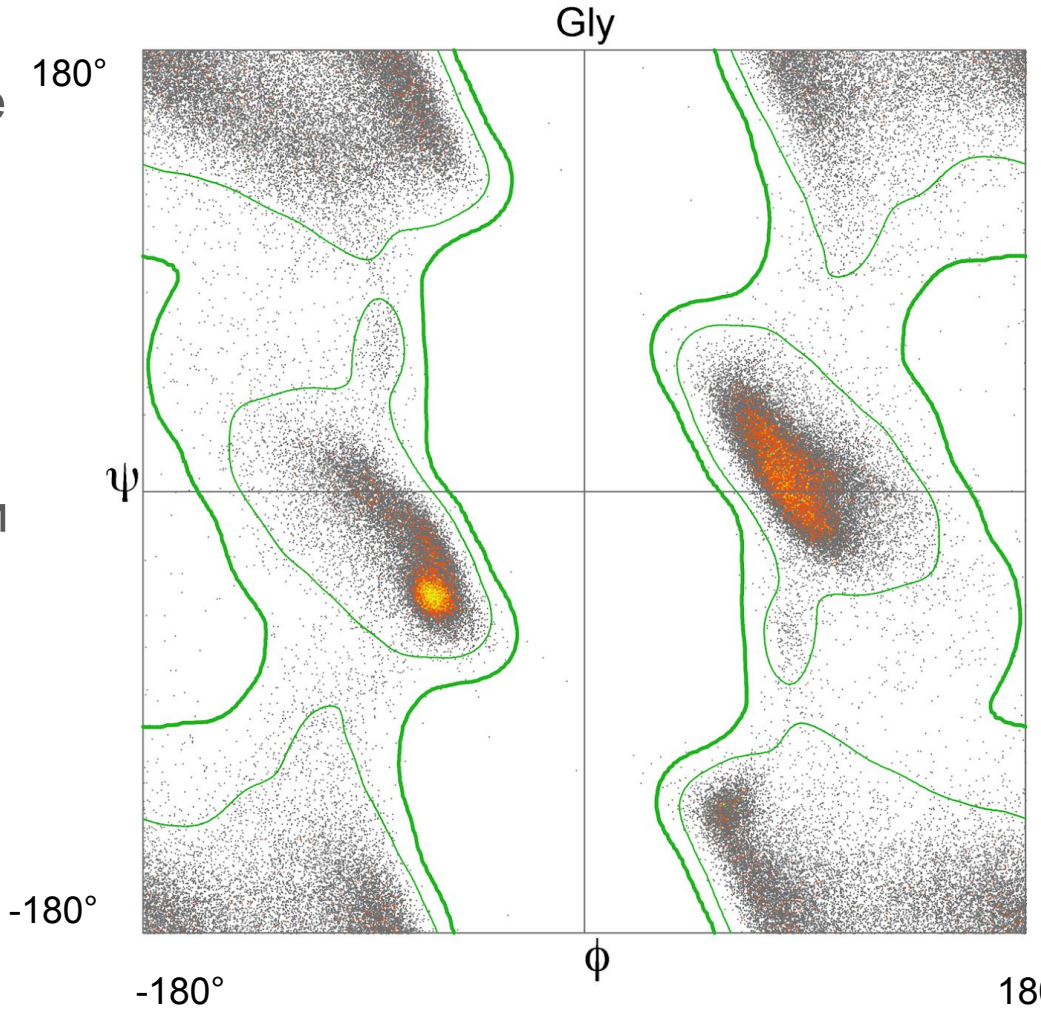
φ и ψ принимают различные значения от -180° до 180°

Показатели качества отдельных остатков: карты Рамачандрана для φ и ψ

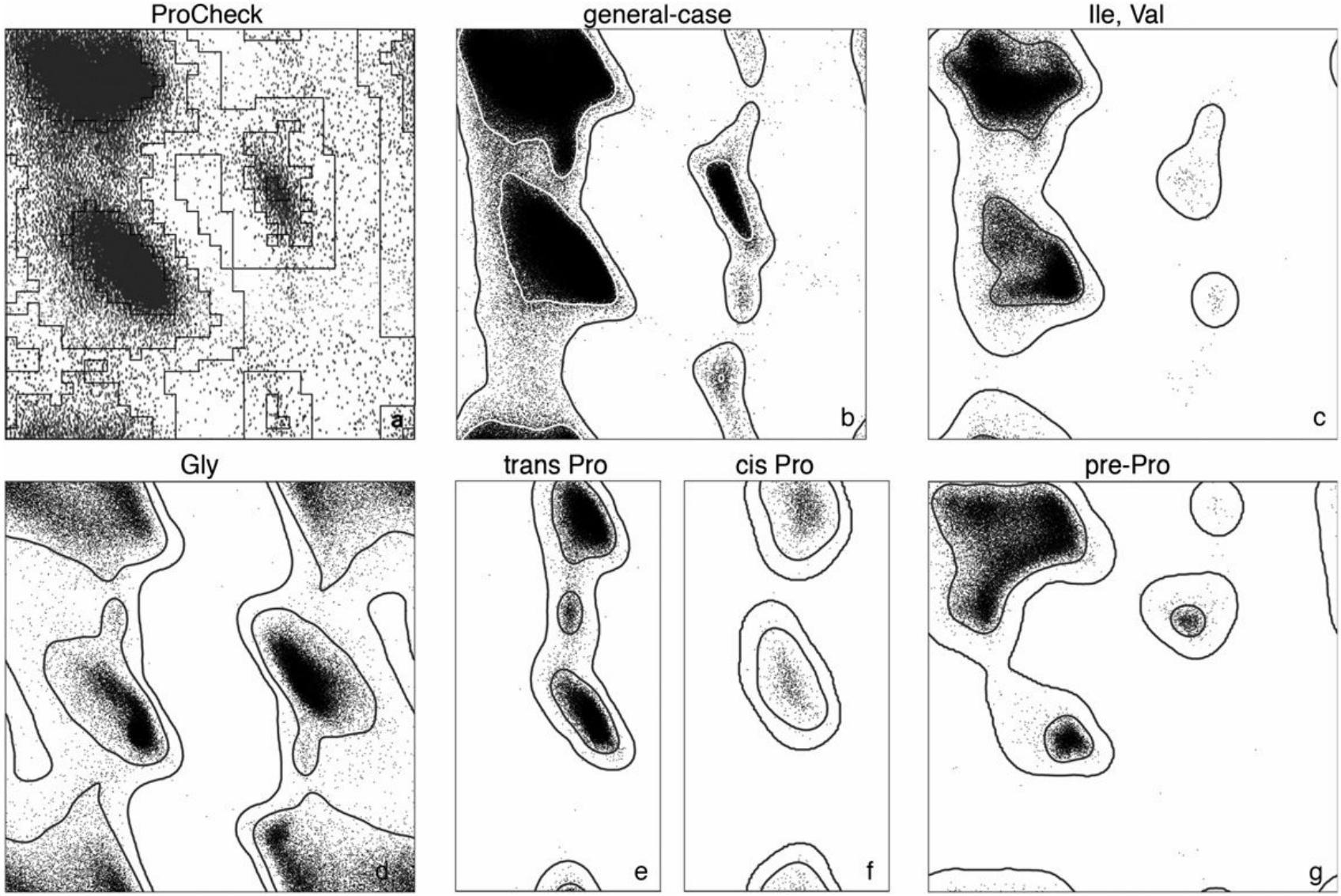
φ и ψ принимают различные значения от -180° до 180°.

Если для каждой аминокислоты на плоскости (φ, ψ) изобразить все наблюдаемые конформации - получится карта Рамачандрана.

Карты для отдельных остатков и для общих случаев строятся на основе статистики базы PDB.



Показатели качества отдельных остатков: карты Рамачандрана для φ и ψ

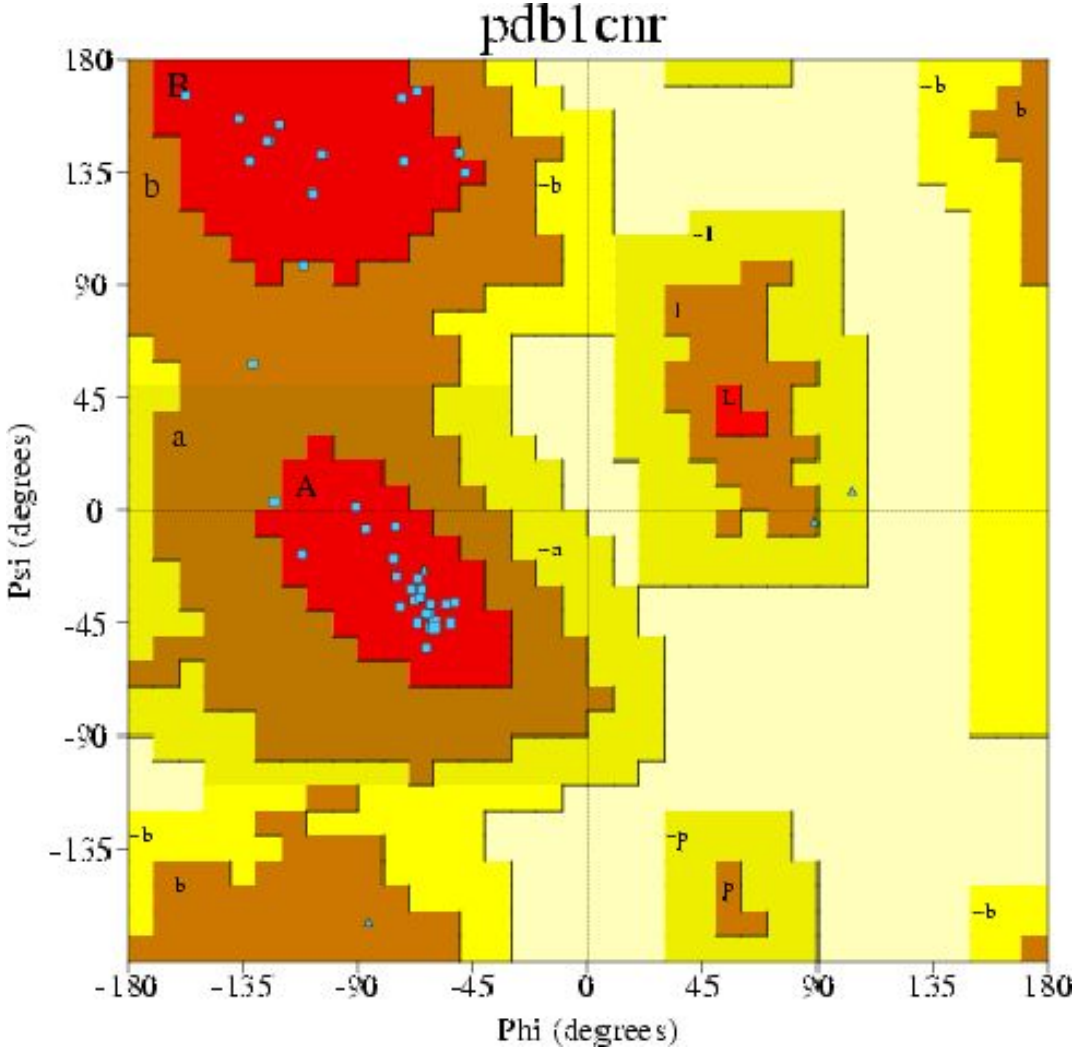


Показатели качества отдельных остатков: карты Рамачандрана для φ и ψ

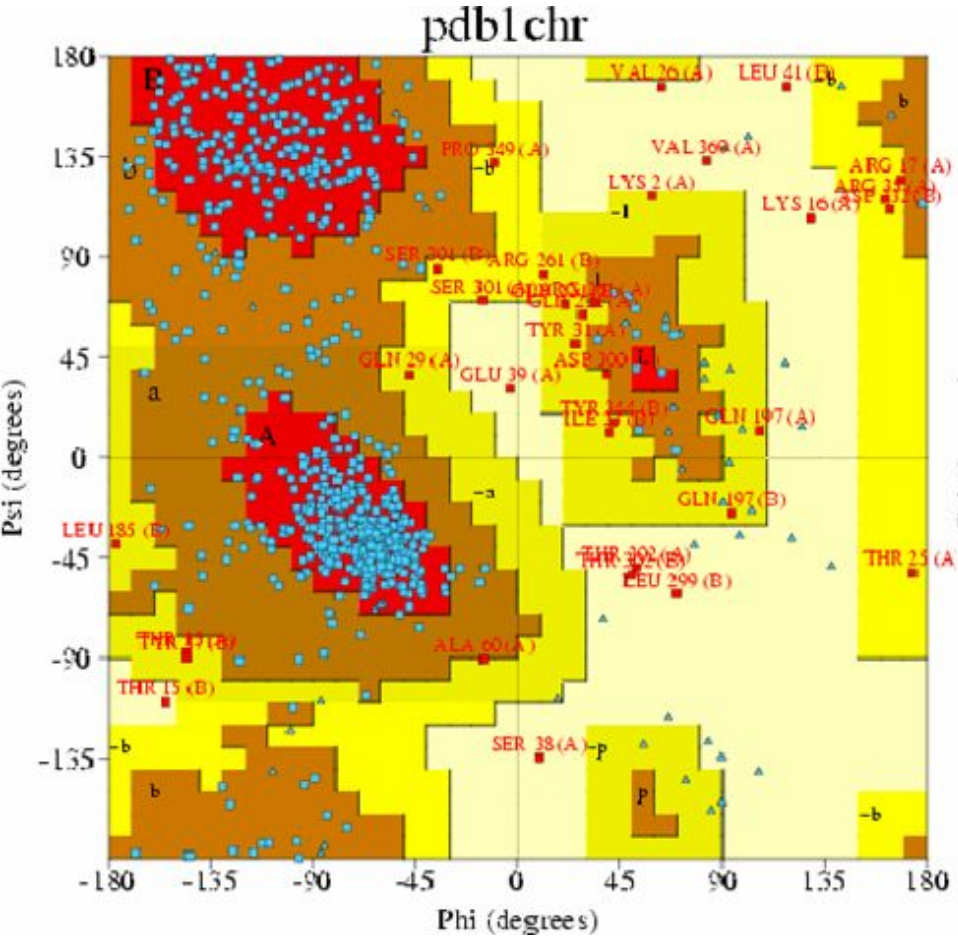
Выделяют предпочитаемую, разрешенную, допустимую, запрещённую области на карте.

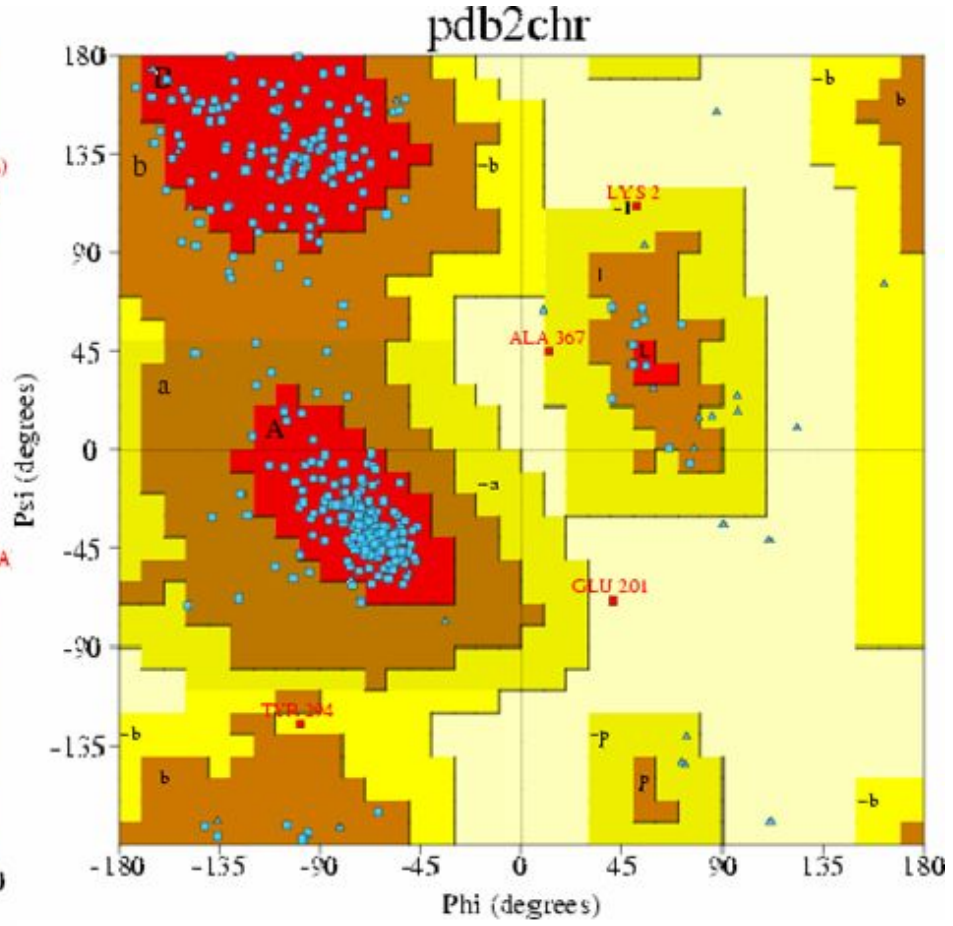
Попадание остатка вне предпочитаемой области - **маргинал.**

В хорошей модели >90% остатков, не считая Gly, Pro, находятся в предпочитаемой области



Показатели качества отдельных остатков: карты Рамачандрана для φ и ψ



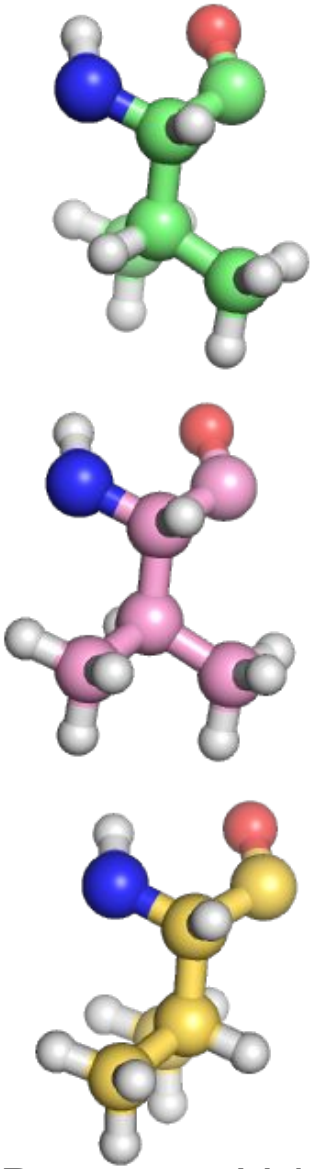


Показатели качества отдельных остатков: торсионы боковых цепей

Боковые цепи имеют от 0 (Gly, Ala) до 4х (Lys, Arg) степеней свободы - вращений вокруг ковалентных связей боковой цепи

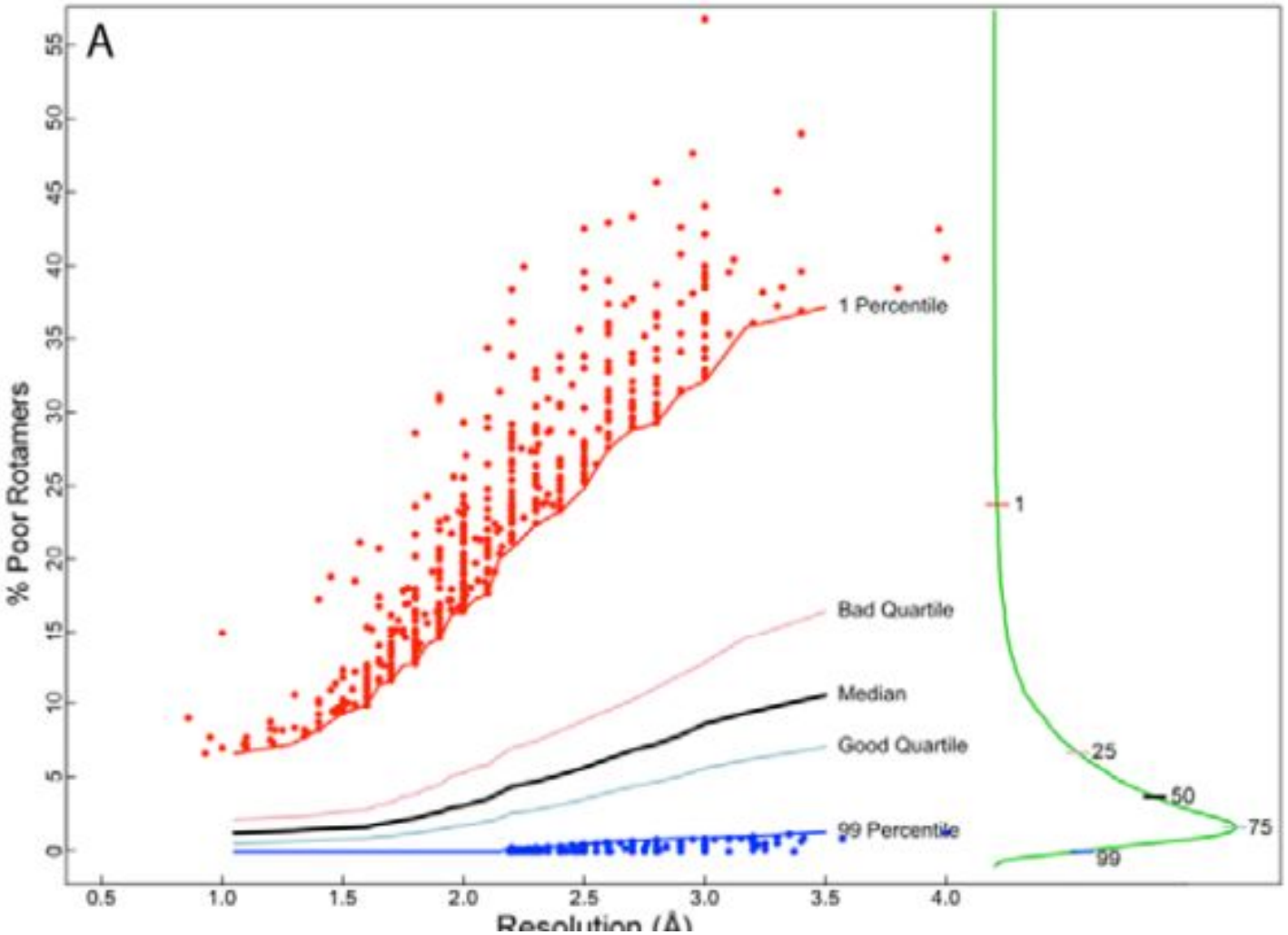
Соответствующие торсионные углы обозначаются χ_1, ..., χ_4, отсчёт идет от связи Cα – Cβ

Ротамеры - типичные конформации боковых цепей остатков. Задаются значениями χ_1, ..., χ_4 и доверительными интервалами к ним. Вычисляются на основе статистики базы PDB.



Ротамеры Val

Показатели качества отдельных остатков: торсионы боковых цепей



Зависимость процента плохих ротамеров от разрешения

Показатели качества отдельных остатков: RSR

Пространственный R-фактор сравнивает соответствие электронной плотности модели и “экспериментальной” электронной плотности. Заменяет визуальный анализ того, как остаток вписан в электронную плотность.

$$RSR = \frac{\sum_{A \in L} |\rho_{\text{эксп}} - \rho_{\text{модель}}|}{\sum_{A \in L} \rho_{\text{эксп}}} [\cdot 100\%]$$

(Суммирование по гриду в пространстве вокруг группы атомов)

Проблема: “экспериментальная” электронная плотность строится **до** оптимизации модели и больше никак не используется. Откуда её взять?

Хорошие значения: RSR**<10%**
Плохие: RSR**>20%**

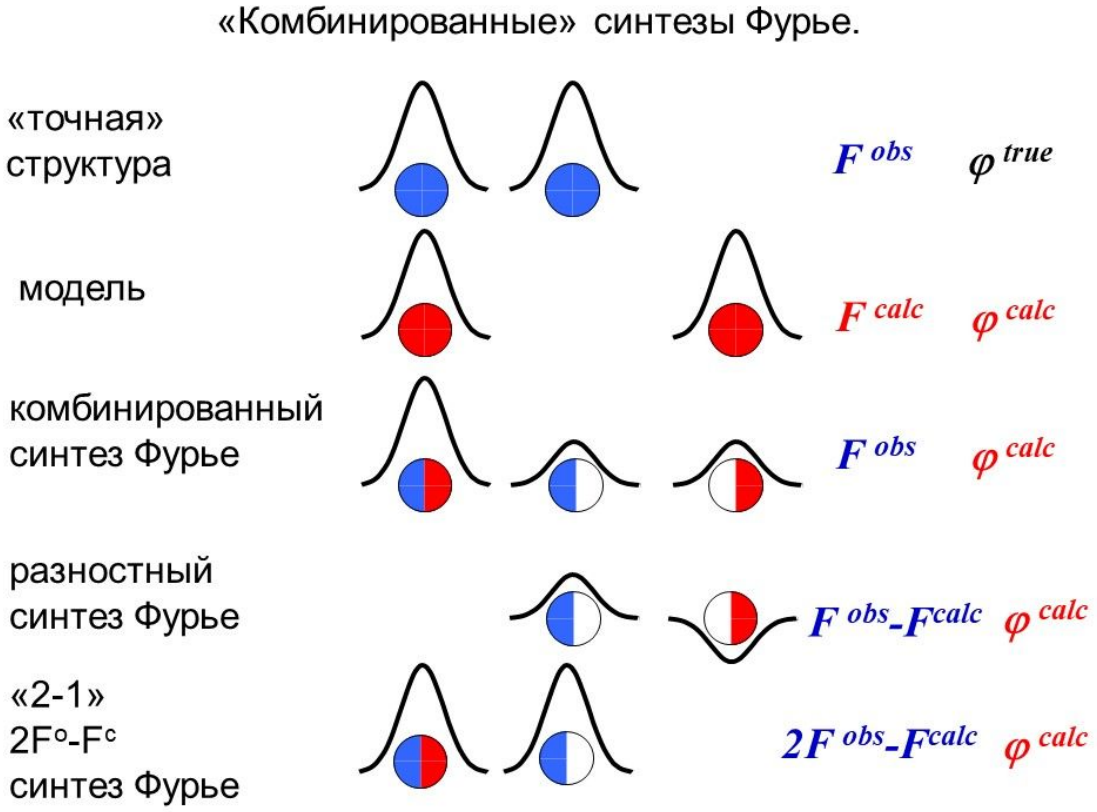
Показатели качества отдельных остатков: RSR

“Экспериментальную” ЭП считают с помощью комбинированного синтеза Фурье: $2F_{эксп}-F_{модель}$, $\varphi_{модель}$

Откуда взять $F_{эксп}$ для структуры в PDB?

С 2008 года при депонировании структуры в PDB авторы обязаны предоставлять файл структурных факторов

Сайт Electron Density Server (EDS) - ЭП для PDB



Показатели качества отдельных остатков: RSR Z-score

Относительная оценка RSR для конкретного остатка.

RSR сравнивается со средним RSR для того же типа остатков (Ala) по выборке из PDB с примерно таким же разрешением

Если RSR плохой, а RSR-Z – хороший, то значит координаты атомов расшифрованы плохо, но не хуже, чем в других подобных структурах.

$$Z = \frac{(RSR - \langle RSR_{resolution} \rangle)}{\sigma(RSR_{resolution})}$$

Хорошие значения: RSR-Z**<2**
Плохие: RSR-Z**>2**

Показатели качества отдельных остатков: “Комфортность” окружения атома

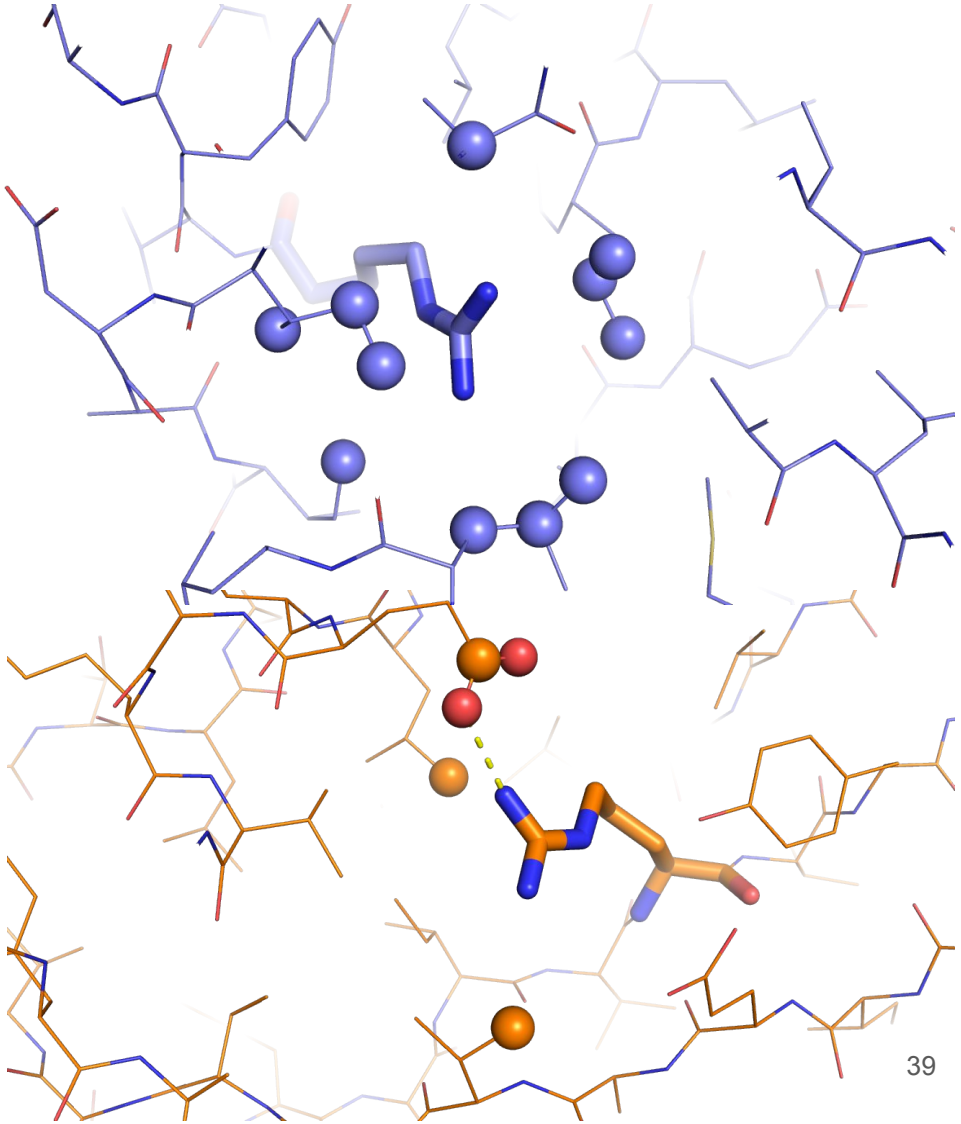
Основывается на физико-химических ограничениях для взаимодействующих групп атомов.

Полностью независим от процедуры оптимизации модели.

Зависим от экспертной оценки.

Показатели качества отдельных остатков: “Комфортность” окружения атома

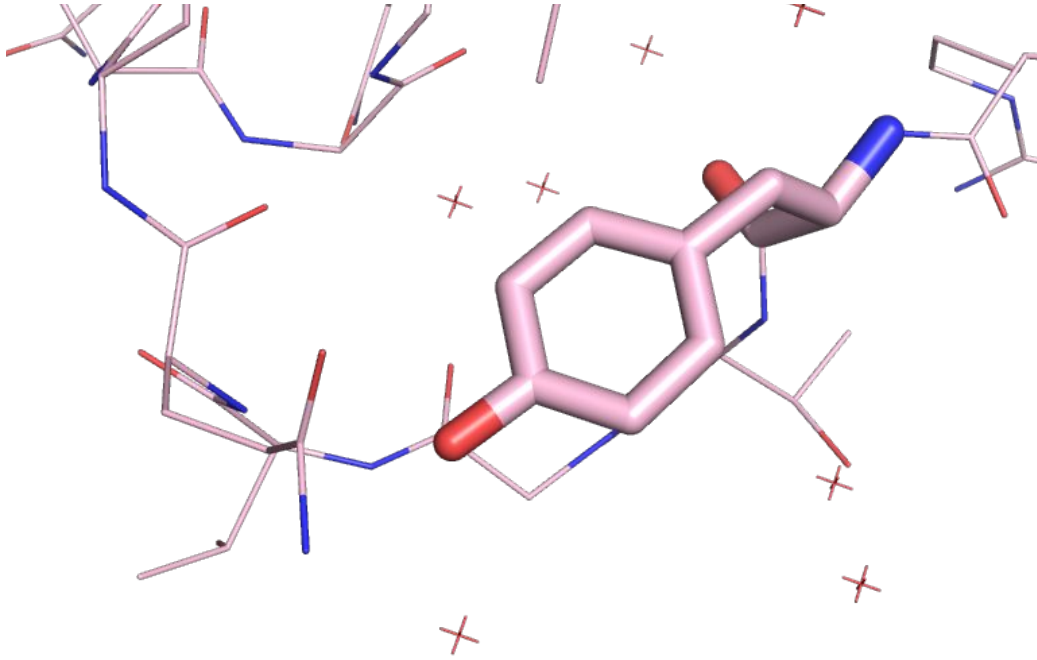
- Clash (перекрывание Ван-дер-Ваальсовых радиусов несвязанных атомов)
- Экспонированные гидрофобные остатки
- Доноры/акцепторы водородной связи без водородной связи
- Заряженные группы не скомпенсированы противоионами



Arg35 в 1chr и 2chr

Показатели качества отдельных остатков: Интегральная оценка комфортности окружения остатка

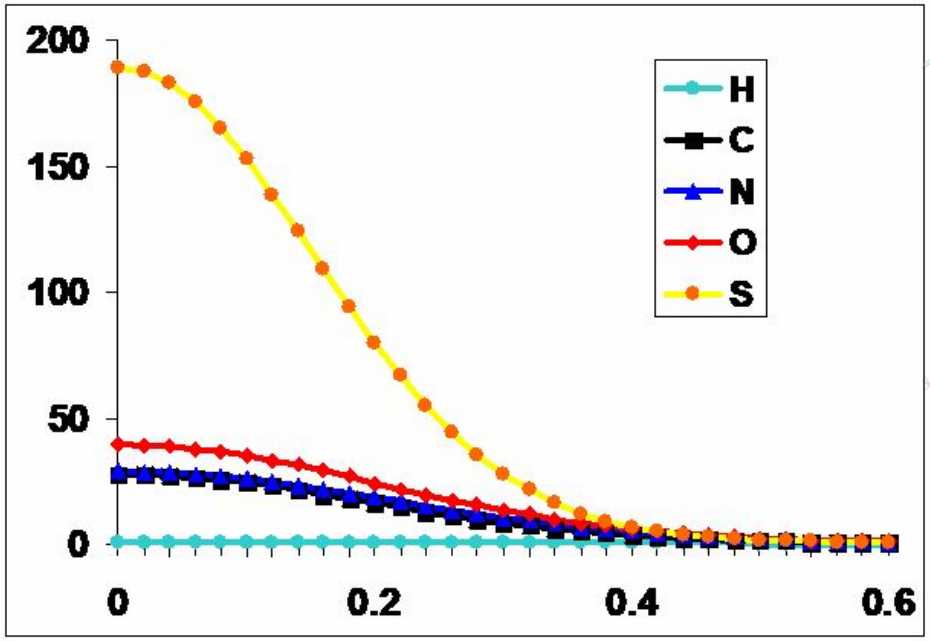
- В программе WhatCheck, например, рассчитывается Z-score для комфортности окружения каждой боковой цепи
- Маргиналы – Z-score < -5
- Маргиналов по окружению стоит проверять визуально: часто маргинальность объясняется выходом на поверхность глобулы, контактом с белком из соседней ячейки и др.

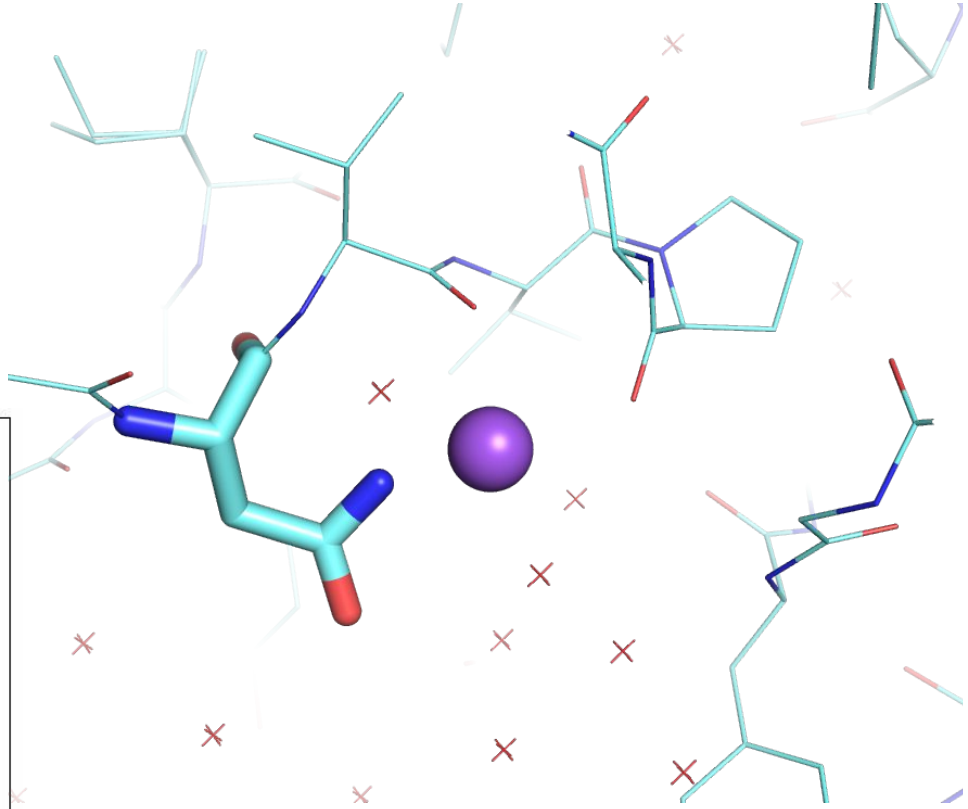


Tyr221 5TLN,
Z-score=5.07

Показатели качества отдельных остатков: Инверсии Asn/Gln/His

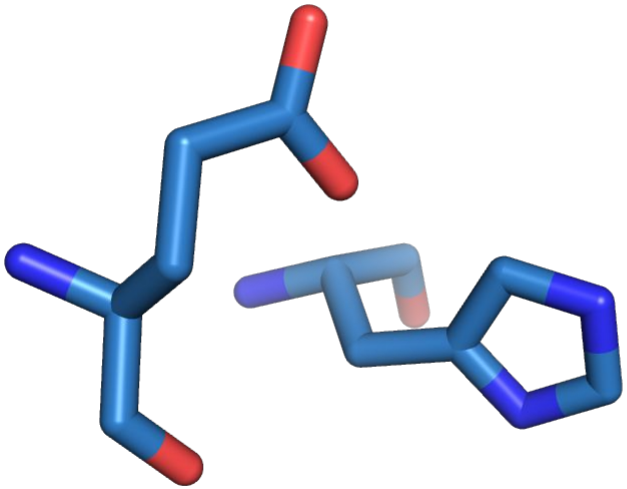
Электронная плотность O и N трудноразличима. Ориентацию Gln/Asn надо проверять.



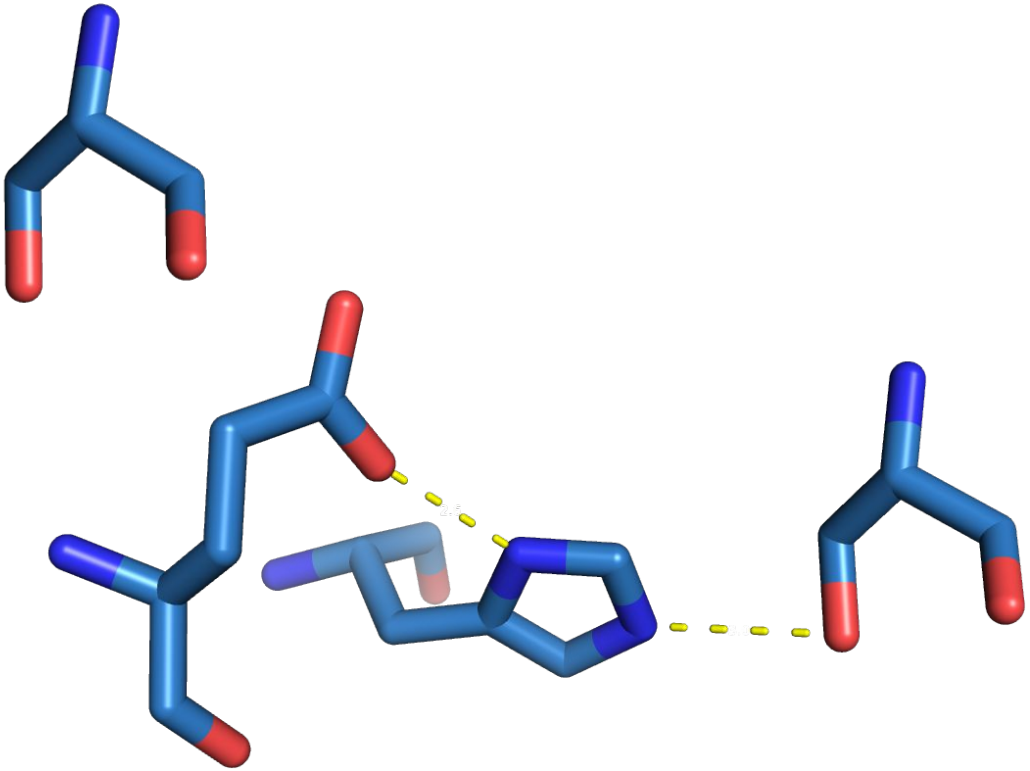


Asn100 1kjq и Na^+

Показатели качества отдельных остатков: Инверсии Asn/Gln/His

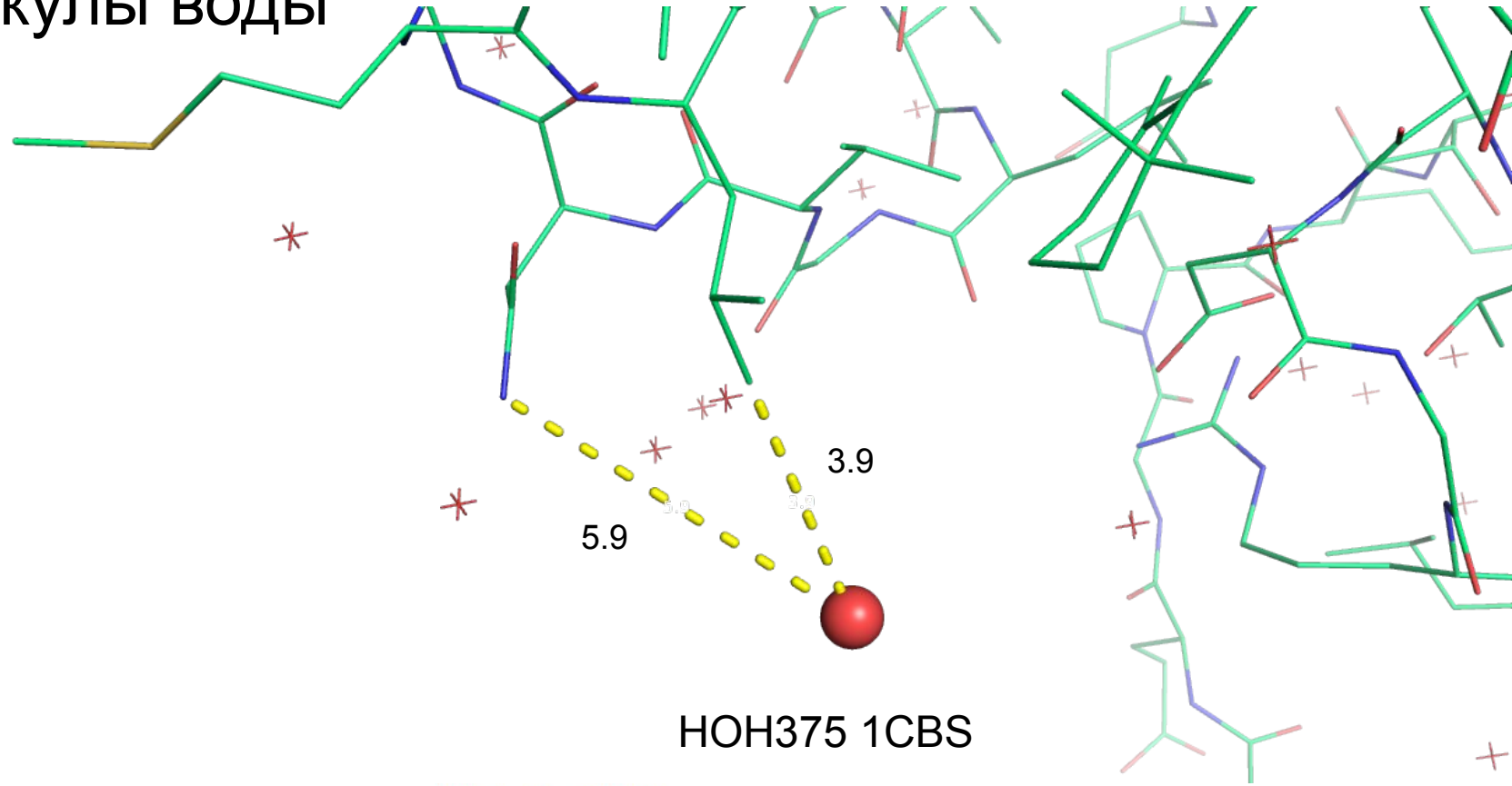


Каталитическая триада в 6ESJ



Функциональная каталитическая триада

Показатели качества отдельных остатков: Молекулы воды



HOH375 1CBS

Не все молекулы воды кристалла не являются ошибочными.

```
REMARK 525 SOLVENT
REMARK 525
REMARK 525 THE SOLVENT MOLECULES HAVE CHAIN IDENTIFIERS THAT
REMARK 525 INDICATE THE POLYMER CHAIN WITH WHICH THEY ARE MOST
REMARK 525 CLOSELY ASSOCIATED. THE REMARK LISTS ALL THE SOLVENT
REMARK 525 MOLECULES WHICH ARE MORE THAN 5A AWAY FROM THE
REMARK 525 NEAREST POLYMER CHAIN (M = MODEL NUMBER;
REMARK 525 RES=RESIDUE NAME; C=CHAIN IDENTIFIER; SSEQ=SEQUENCE
REMARK 525 NUMBER; I=INSERTION CODE):
REMARK 525
REMARK 525  M RES CSSEQI
REMARK 525    HOH A 376        DISTANCE =  5.94 ANGSTROMS
```

Remediation - повторное построение моделей

A major focus of the wwPDB is maintaining consistency and accuracy across the archive. As the PDB grows, developments in structure determination methods and technologies can challenge how all structures are represented. The wwPDB addresses these challenges with regular reviews of data processing procedures and coordinates remediation efforts to improve data representation.