Структурная биоинформатика

Практикум 6. Валидация.

Поддъяков Иван

Оглавление

Задание 1.

В данном практикуме я буду работать со структурой 1JZ7 [1]. Это структура фермента β-галактозидазы *E.coli* со связанным лигандом - β-D-галактопиранозой, полученная в рамках исследования по установлению каталитического механизма [2]. Также в структуре присутствуют DMSO и ионы магния и натрия. Ассиметрическая единица представляет собой гетеротетрамер. Отчет валидации wwPDB.

Структура получена с разрешением 1.50 Å при покрытии интервала рефлексов 40.00 - 1.50 Å 92.0% - по данным авторов и также с разрешением 1.50 Å при покрытии интервала рефлексов 35.20 - 1.50 Å 90.9% - по данным валидации, что кажется мне не самым лучшим результатом. Температура проведения эксперимента - 100 К.

В структуре отсутствуют атомы с нулевой заселенностью, для 1 аминокислотного остатка в каждой цепи показана альтернативная конформация. Цистеин-247 заменен на S-гидроксицистеин - именно для его атома OD показано 2 альтернативных положения с равной заселенностью.

По результатам валидации значение R-фактора для модели 0.169 и R_{free} 0.207, что говорит о достаточно хорошем соответствии модели экспериментально полученным моделям структурных факторов; R_{free} - R < 10%, скорее всего модель не переоптимизирована.

В среднем в модели по данным валидации 0.6% аминокислотных остатков являются маргинальными по длине связи, 1.7% являются маргинальными по углам связи. Для цепей значения значительно не различаются.

В модели определены 530 слишком близких контактов (клэшей) с учетом добавленных водородов, в среднем 8 клэшей на 1000 атомов.

Для остова 1008 проанализированных остатков на цепь в среднем предсказано 0% (12 из 4032) маргиналов по углам Ψ и Φ, при этом 4 % отнесены к категории допустимых.

Относительно типичных торсионных углов боковых цепей (ротамеров) маргинальными являются 4% остатков при проанализированных 863 на цепь, что является достаточно плохим результатом.

RSR Z-score выше 2 только у 2% аминокислотных остатков, что лучше, чем соответствие модели и измеренных рефлексов у двух третей других структур. Данных по RSR для аминокислот в отчете по валидации не представлено.

На Рисунке 1 показано соотношение между метриками данной модели и всех других в банке PDB. Выделяется относительно очень высокий процент маргинальных

остатков по положению боковых цепей - только у 8% структур хуже. Также показатели по количеству клэшей и маргиналов по углам остова ниже среднего. R_{free} и RSRZ у данной модели около среднего, что может говорить о неплохом соответствии модели и экспериментальных данных. Однако все показатели сильно смещаются в худшую сторону, если рассматривать структуры со схожим разрешением, что, наряду с покрытием интервала разрешений порядка 90%, вызывает у меня сомнения в информативности вклада полученных рефлексов с маленькими d в общую модель.

На Рисунке 2 представлено сопоставление качества разных цепей. Зеленым показан процент остатков, не являющихся маргиналами по геометрическим критериям, желтым - являющимися по одному критерию, оранжевым - по 2, красным - по 1. Серым указана доля остатков, не представленных в модели (потери при клонировании). Красным сверху - процент остатков, плохо вписывающихся в свою ЭП. Видно, что цепи сильно похожи между собой по этим параметрам. Далее я буду работать преимущественно с цепью A, если не указано иного.

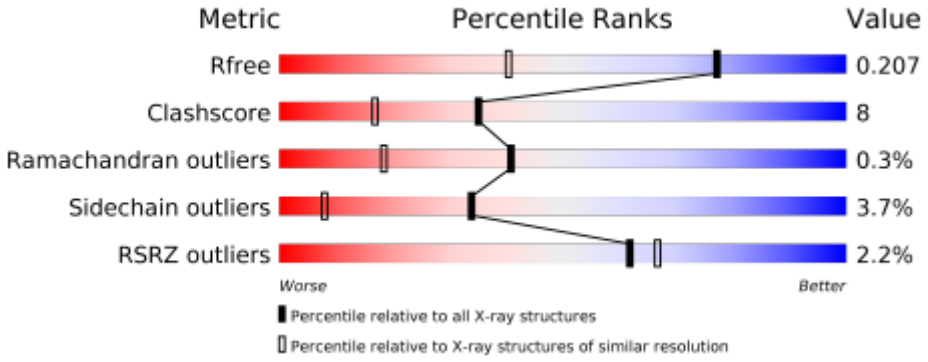

Рисунок 1. Параметры качества модели относительно банка PDB.

Mol	Chain	Length	Quality of chain
1	A	1023	2% 71% 24%
1	B	1023	2% 72% 22%
1	C	1023	2% 71% 22% 5%
1	D	1023	3% 70% 24% 5%

Рисунок 2. Качество белковых цепей в структуре.

Таблица 1. Остатки с аннотированной функцией и их расшифровка.

Ам. к-та	№ в UniProt/PDB	Аннотированная функция	Отчет валидации wwPDB	Отчет MolProbity после исправления
N	103/102	Связывание субстрата	Клэш	Аналог.
D	202/201	Связывание субстрата	Маргинал по углам	Аналог.
H	358/357	Стабилизация ПС	Клэш	Клэш, маргинал по углу
H	392/391	Стабилизация ПС	Минимальный клэш	Маргинал по длине связи
E	417/416	Связывание магния	Клэш, маргинал по длине связи и углам	Маргинал по углам, клэш
H	419/418	Связывание магния	Клэш	Клэш, маргинал по длине связи и углам
E	462/461	Донор протона в активном центре	Маргинал по длине связи	Аналог.
E	538/537	Нуклеофил в активном центре	2 клэша, маргинал по длине связи	Аналог.
N	598/597	Связывание магния	-	Маргинал по углу
F	602/601	Связывание натрия	Клэш	Клэш, маргинал по углу
N	605/604	Связывание субстрата	Минимальный клэш	Аналог.
W	1000/999	Регуляция	Клэш	Аналог.

В структуре имеются участки, которые плохо вписываются в ЭП и для которых ее значения существенно более размыты. Я убедился, что функциональные остатки хорошо подтверждаются ЭП (Рисунок 3).

В Таблице 1 описаны ошибки в моделировании остатков, которые имеют аннотированную функцию в Uniprot, по данным отчета валидации wwPDB и MolProbity после разворота позиций и добавления водородов. Примечательно, что все функциональные остатки являются маргинальными по одному или нескольким критериям

качества, как наличие клэшей или сильно отклоняющихся углов и длин связей. Возможно, атомы этих остатков были вписаны таким образом в процессе оптимизации, потому что обладают подвижностью, свойственной остаткам активного центра, и их экспериментальная ЭП немного размыта, отражая их различное положение при связывании лиганда.

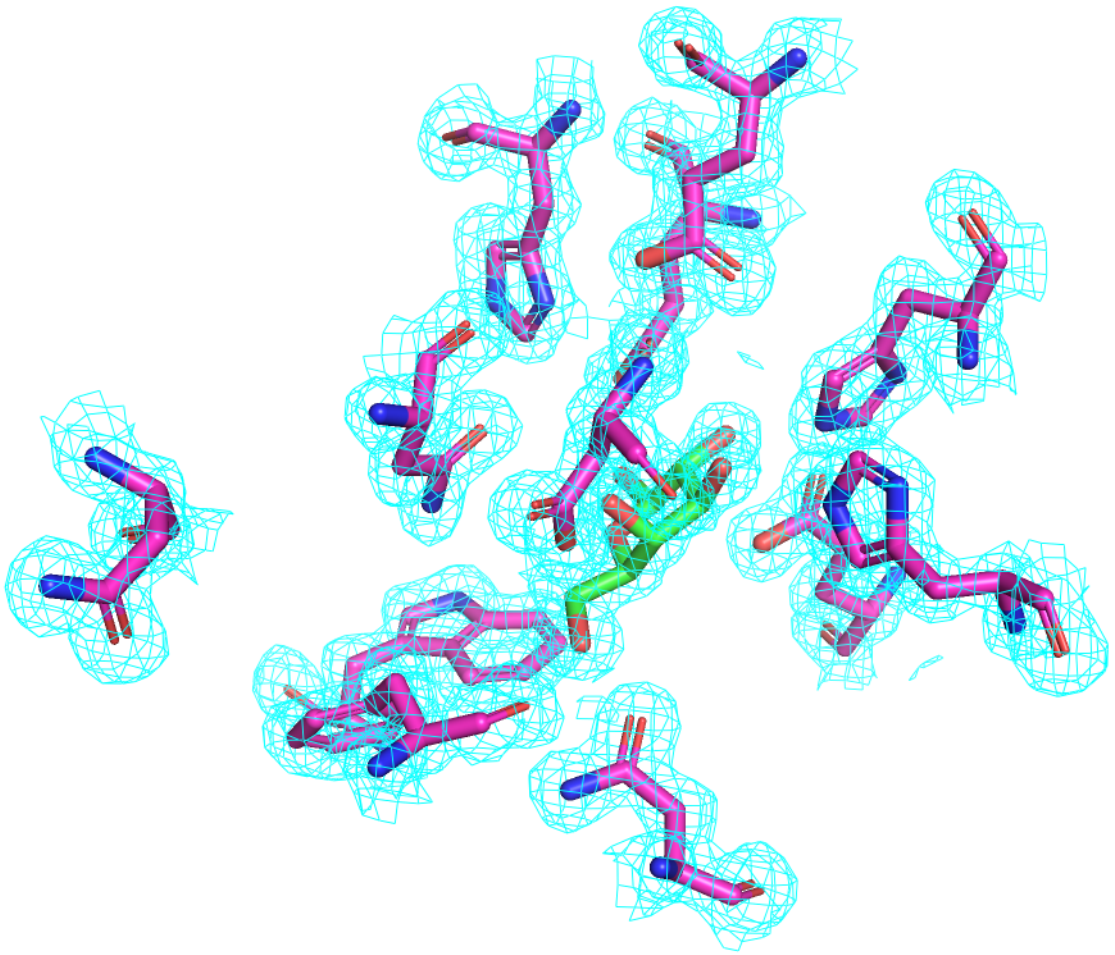

Рисунок 3. Функциональные участки (фиолетовый), лиганд GAL (зеленый) и их электронная плотность на уровне подрезки 2.

Задание 2.

2.1. Маргинальные остатки.

В данном задании я выбрал некоторые маргинальные остатки из отчетов wwPDB и MolProbity и визуализировал их.

На Рисунке 4 представлен гистидин-735, который имеет самый высокий RSRZ-score 8.4 во всей модели по данным отчета wwPDB. Видно, что даже на уровне подрезки 0.7, ЭП очень плохо накладывается на остов и боковую цепь данного остатка. Это может объясняться тем, что данный остаток находится на самой поверхности белковой молекулы и как подвижен в растворе, так и подвержен тепловым колебаниям в кристалле. Некоторые его атомы имеют B-фактор равный 100 $Å^2$.

Сервис MolProbity, помимо прочего, позволяет найти аминокислотные остатки (HIS, GLN), боковую группу которых следует развернуть, чтобы устранить клэши водородов с соседями. На Рисунке 5 как раз показан такой случай, где водород на аминогруппе боковой цепи глутамина находится слишком близко к водороду остова (перекрытие 0.82 Å). На Рисунке 6 изображен тот же остаток после обработки на сервере MolProbity. Клэша больше нет, но остаток также хорошо вписывается в электронную плотность - скорее всего, получена наиболее достоверная конформация этой боковой цепи.

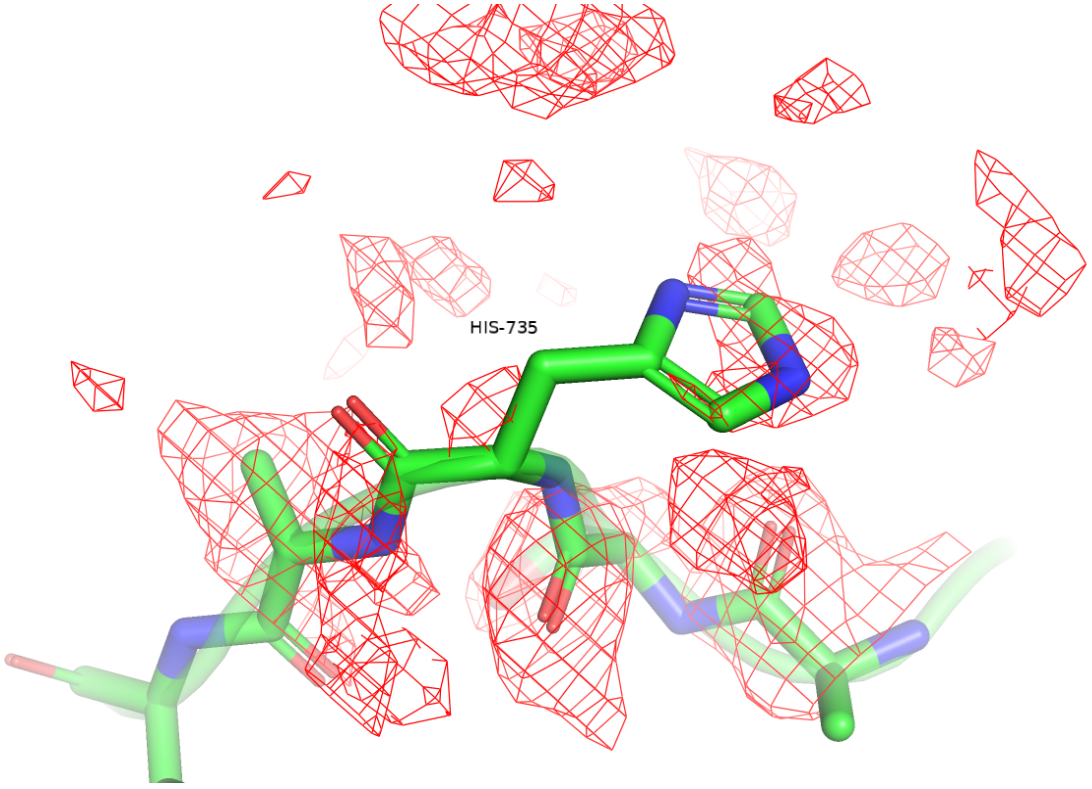

Рисунок 4. Гистидин-735, вписанный в ЭП, уровень подрезки 0.7.

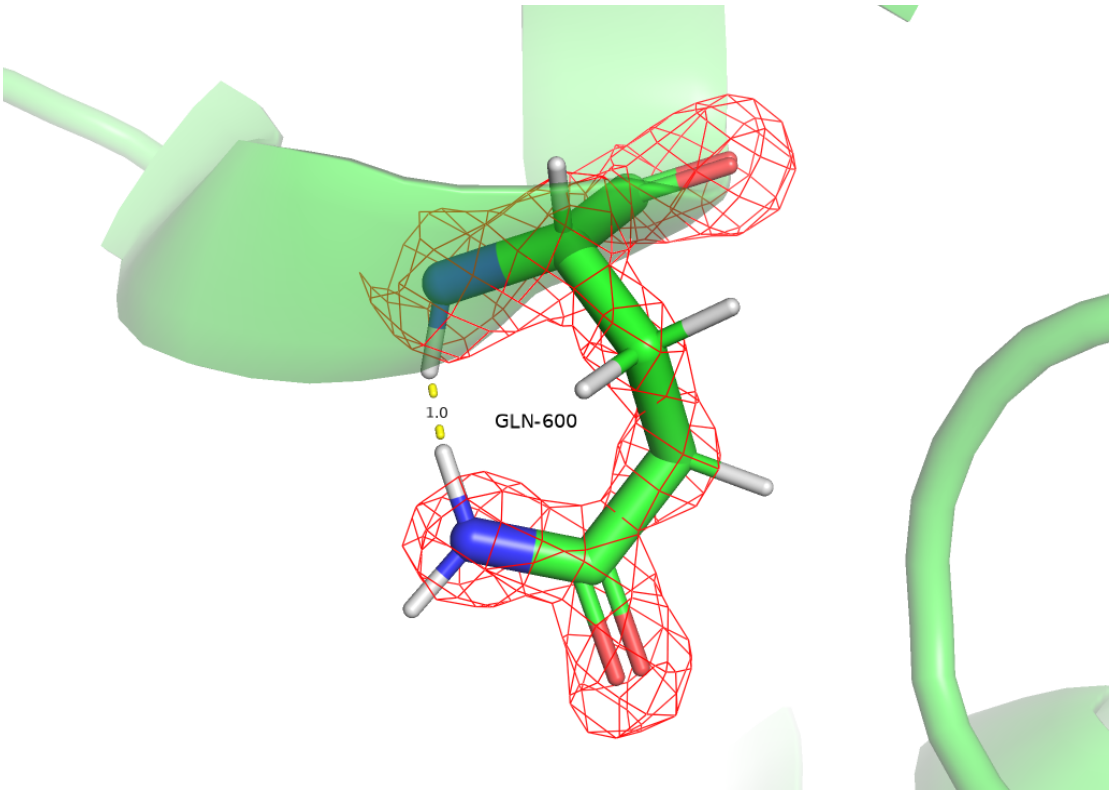

Рисунок 5. Глутамин-600, вписанный в ЭП, с клэшем между водородами, уровень подрезки 2.

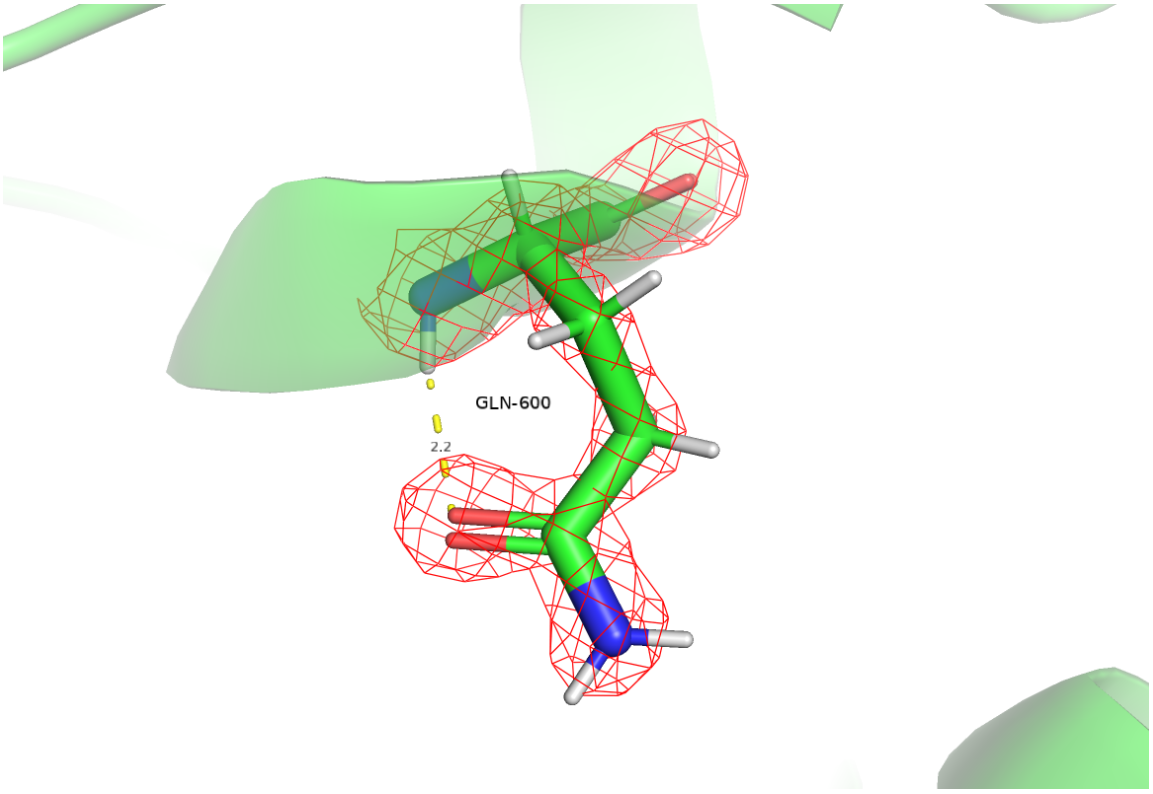

Рисунок 6. Глутамин-600, вписанный в ЭП (уровень подрезки 2), боковая группа перевернута MolProbity.

Остатки также могут быть признаны маргинальными, если торсионные углы боковых цепей сильно отличаются от типичных значений. На Рисунке 7 показан остаток тирозина-538, в модели которого торсионные углы несколько (на 5 и более стандартных отклонений) отличаются от идеальных. Угол CZ-CE2-CD2 отклоняется на 4.7 градуса, угол CE2-CD2-CG больше на 5,2 градуса, угол CD2-CG-CB больше на 3,2 градуса, чем ожидаемое значение.

На Рисунках 8 и 9 представлено разрешение другого клэша между соседними атомами с перекрытием 0.44 Å. После инверсии имидазольной группы расстояние между накладывающимися CE1 гистидина и HB3 серина увеличилось.

Еще один случай клэша приведен на Рисунке 10, перекрытие 0.43 Å. Кислород OE2 глутамата-537 очень сближен с водородом H1 галактопиранозы, при этом и связь CD-OE2 длиннее, чем ожидается, на 0.09 Å, что также позволяет отнести остаток к маргиналам. Возможно, корректировка положения данного кислорода приведет к разрешению этих двух неточностей сразу.

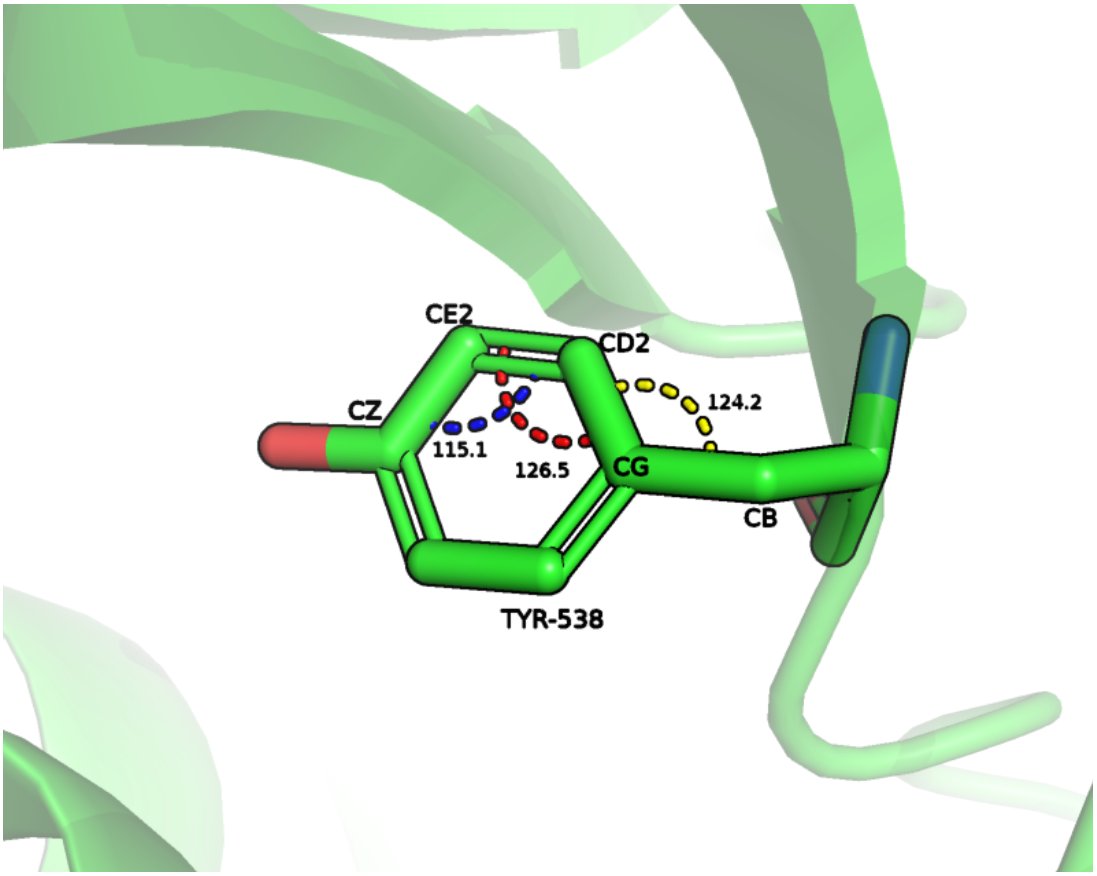

Рисунок 7. Тирозин-538 и значения углов.

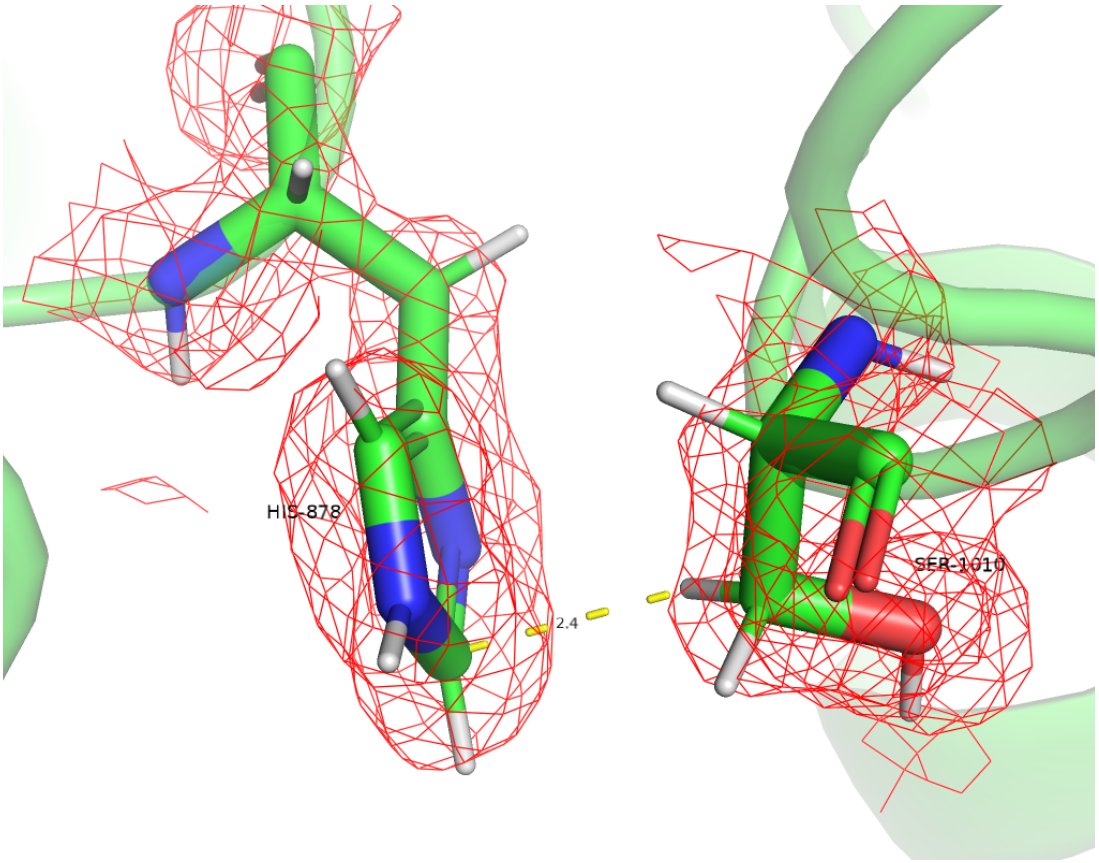

Рисунок 8. Клэш гистидина 878 и серина-1010.

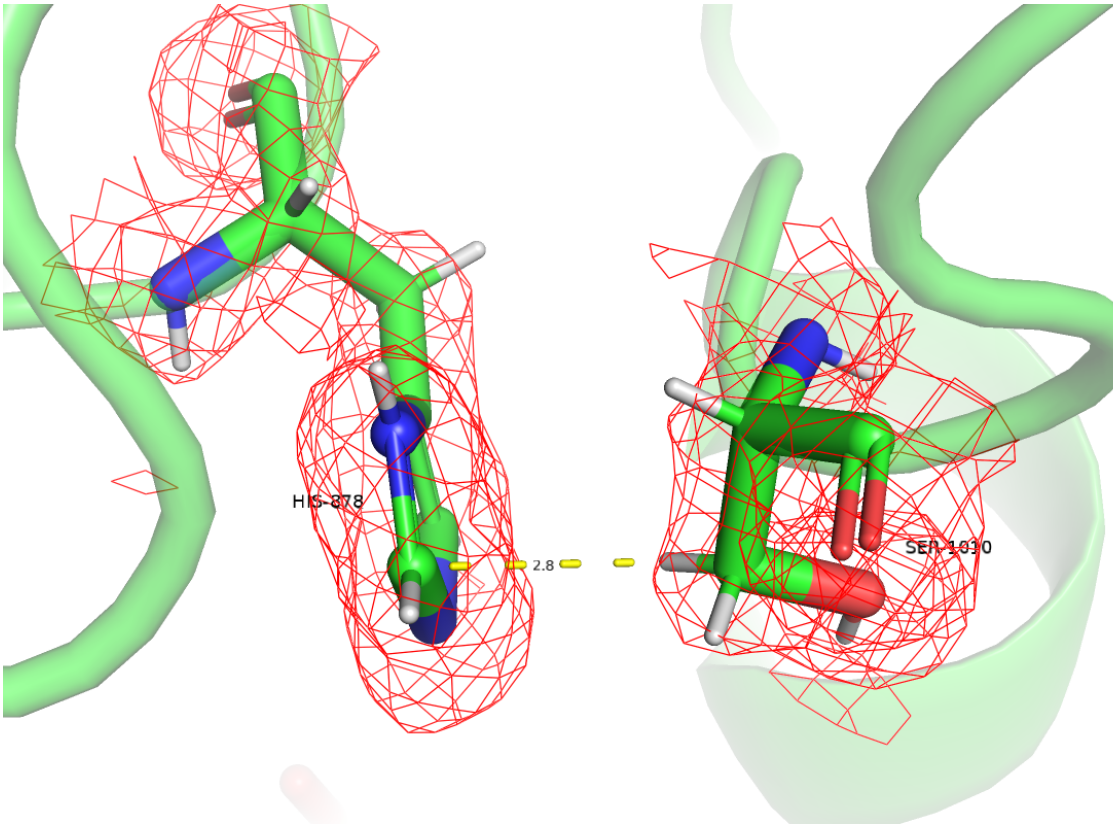

Рисунок 9. Гистидин-878 перевернут MolProbity, клэша нет.

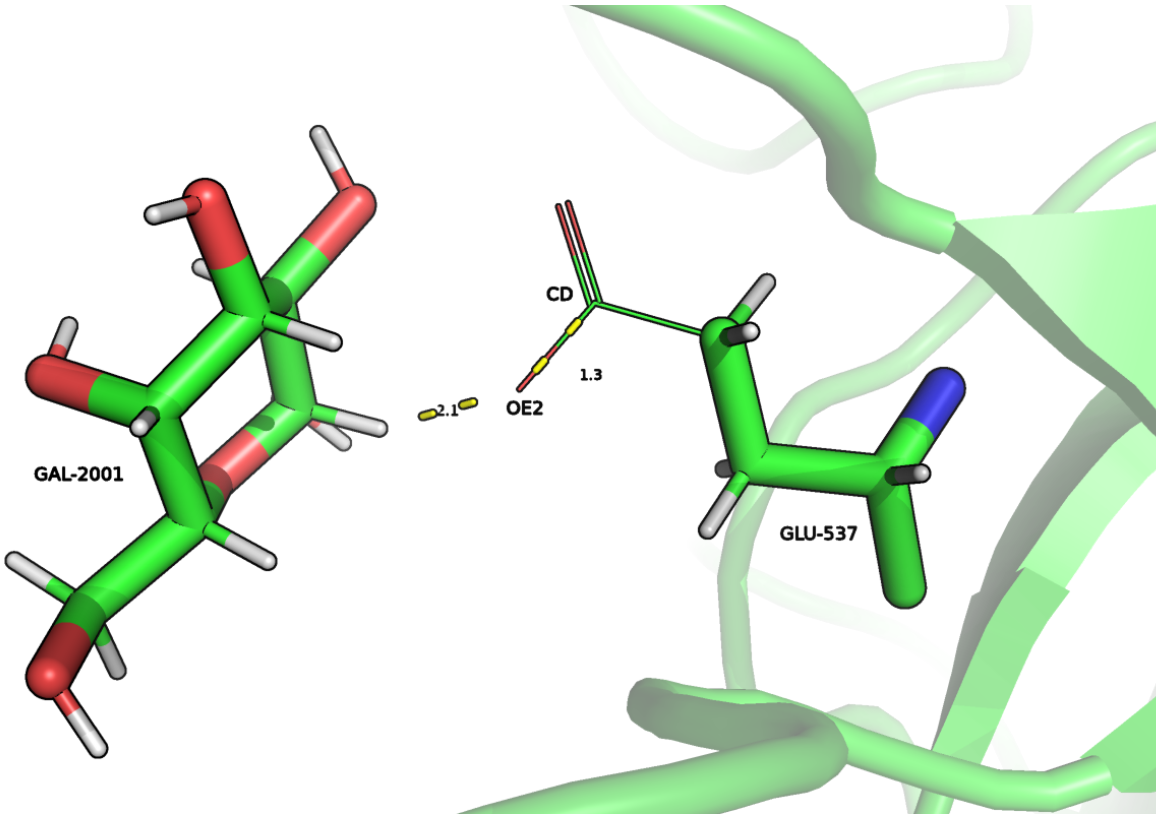

Рисунок 10. Глутамат-537 и лиганд GAL.

2.2. Отчеты валидации.

MolProbity выдает схожие с wwPDB показатели качества структуры и остатков, иногда добавляя новые остатки в маргинальные так как, видимо, имеет более низкий порог отсечения по Z-оценке. На Рисунке 11 представлена общая информация о качестве структуры по данным MolProbity без дополнительного улучшения. Видно, что этот сервис определяет больше (367 против 260) маргиналов по длине связи, больше (966 против 808) маргиналов по торсионному углу боковой цепи. Оба отчета находят 12 остатков, которые являются выбросами по торсионным углам остова.

После того, как я добавил водороды и MolProbity предложил и произвел 46 инверсий, в сводной таблице по всем контактам улучшилась метрика наложений (Clashscore) на 0.7 и углы 1 остатка стали лучше соответствовать ротамерам (Рисунок 12).

По выдаче CheckMyMetal см. 4.2. Отчеты валидации.

All-Atom Contacts	Clashscore, all atoms:	8.84		68th percentile* (N=598, 1.50Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	131	3.79%	Goal: <0.3%
	Favored rotamers	3131	90.70%	Goal: >98%
	Ramachandran outliers	12	0.30%	Goal: <0.05%
	Ramachandran favored	3871	96.20%	Goal: >98%
	Rama distribution Z-score	-0.81 ± 0.12		Goal: abs(Z score) < 2
	MolProbity score^	2.17		28th percentile* (N=4836, 1.50Å ± 0.25Å)
	Cβ deviations >0.25Å	333	8.86%	Goal: 0
	Bad bonds:	367 / 33845	1.08%	Goal: 0%
	Bad angles:	966 / 46077	2.10%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	20 / 248	8.06%	Expected: ≤1 per chain, or ≤5%
	Cis nonProlines:	12 / 3792	0.32%	Goal: <0.05%
Additional validations	Tetrahedral geometry outliers	14		
	Waters with clashes	211/4214	5.01%	See UnDowser table for details

Рисунок 11. Сводная таблица MolProbity до добавления водородов.

All-Atom Contacts	Clashscore, all atoms:	8.14		74th percentile* (N=598, 1.50Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	130	3.77%	Goal: <0.3%
	Favored rotamers	3131	90.70%	Goal: >98%
	Ramachandran outliers	12	0.30%	Goal: <0.05%
	Ramachandran favored	3871	96.20%	Goal: >98%
	Rama distribution Z-score	-0.81 ± 0.12		Goal: abs(Z score) < 2
	MolProbity score^	2.14		30th percentile* (N=4836, 1.50Å ± 0.25Å)
	Cβ deviations >0.25Å	333	8.86%	Goal: 0
	Bad bonds:	367 / 33845	1.08%	Goal: 0%
	Bad angles:	968 / 46077	2.10%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	20 / 248	8.06%	Expected: ≤1 per chain, or ≤5%
	Cis nonProlines:	12 / 3792	0.32%	Goal: <0.05%
Additional validations	Tetrahedral geometry outliers	14		
	Waters with clashes	234/4214	5.55%	See UnDowser table for details

Рисунок 12. Сводная таблица MolProbity после внесения инверсий.

Задание 3.

Если предполагать использование данной структуры для изучения особенностей этого белка, то я бы задумался о поиске другой модели этого фермента с более хорошими метриками качества. Соответствие модели и экспериментальной электронной плотности хорошее, но качество электронной плотности на высоких разрешениях кажется мне сомнительным, ввиду низкого покрытия. В модели определяется слишком большой процент маргиналов по ротамерам, типичным углам связи и длинам связей. Все остатки с аннотированной функцией являются маргинальными, как показано в Таблице 1.

Задание 4.

4.1 Маргинальные остатки.

В данном задании я воспользовался сервисом PDB Redo, чтобы переделать модель и улучшить ее качество. Сначала рассмотрим изменения в маргинальных остатках из задания 2. Для некоторых из них удалось достичь качественного улучшения в модели.

На Рисунке 13 изображен глутамат-537 и лиганд галактопираноза. В этом случае в переделанной модели осталось нетипичное расстояние между CD и OE1 глутамата, но и наложение OE1 на водород GLA стало более сильным - перекрытие возросло на 0.13 Å. В данном случае сервис PDB Redo не улучшил модель.

На Рисунке 14 представлен остаток тирозина-538. В переделанной структуре маргинальные углы стали лучше соответствовать типичным значениями, отличаясь от них не более чем на 1 градус.

На Рисунке 15 показан глутамин-600, в котором в исходной модели водород на аминогруппе боковой цепи глутамина находился слишком близко к водороду остова. После перестроения модели PDB Redo перевернул боковую группу глутамина, устранив изначальное наложение почти также хорошо, как MolProbity. Описанный клэш, тем не менее, остался в цепи B.

На рисунке 16 показаны соседние гистидин-878 и серин-1010. В изначальной структуре атомы CE1 гистидина и HB3 серина “сталкивались” друг с другом. PDB Redo перевернул имидазольную группу гистидина и устранил этот клэш.

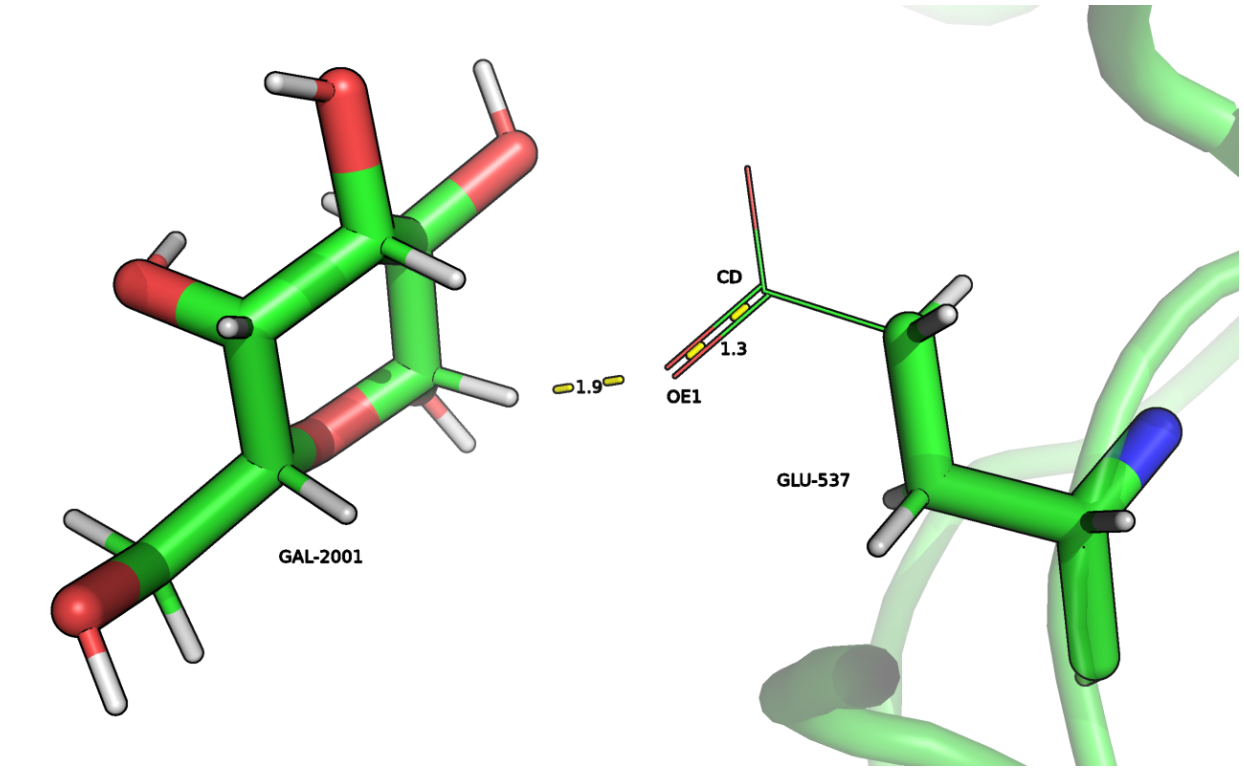

Рисунок 13. Глутамат-537 и лиганд GAL, PDB Redo.

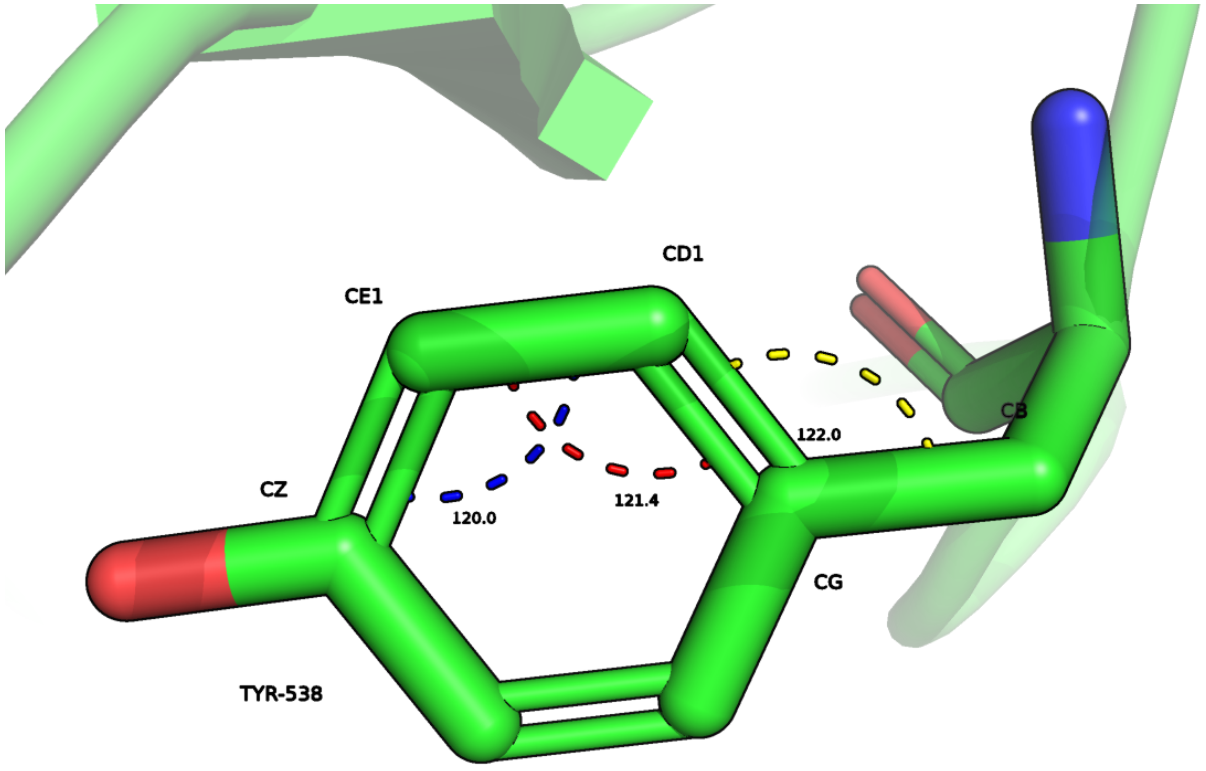

Рисунок 14. Тирозин-538 и значения углов, PDB Redo.

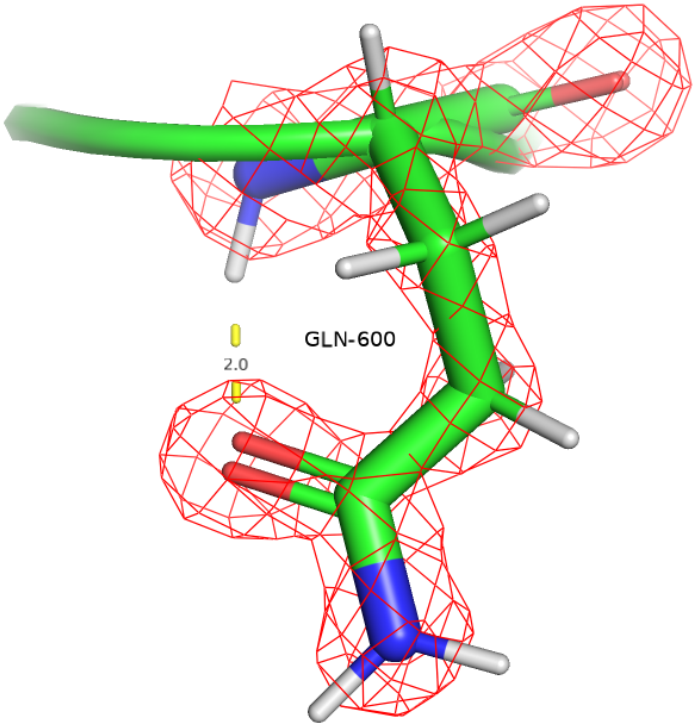

Рисунок 15. Глутамин-600, вписанный в ЭП (уровень подрезки 2), боковая группа перевернута PDB-Redo.

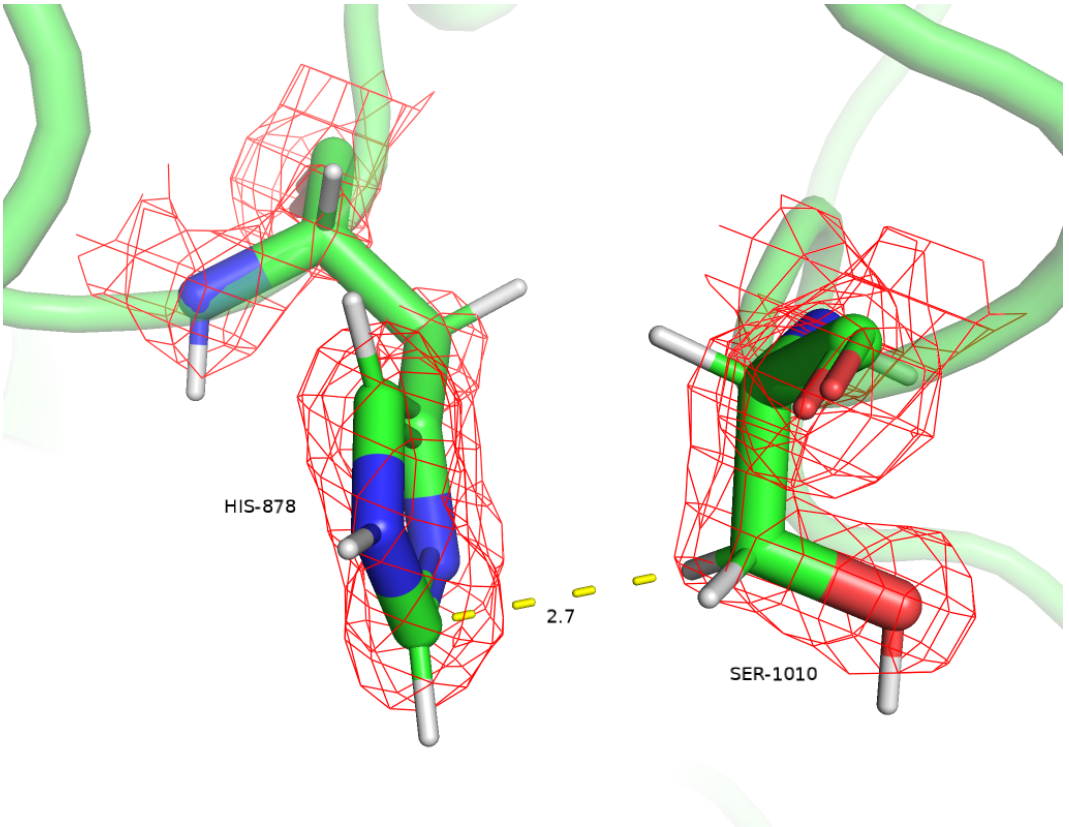

Рисунок 16. Гистидин-878 перевернут PDB-Redo, клэша с серином-1010 нет.

Рассматривая, как измененная модель “укладывается” в экспериментальную ЭП, я нашел любопытный случай. на Рисунке 17 представлен остаток гистидина-646 в исходной и измененной структурах с отображением ЭП на трех уровнях подрезки. При этом в изначальной структуре на месте смещенного Redo имидазольного кольца находится молекула воды, с которой изначальный гистидин в цепи A имеет незначительный клэш. ЭП на уровне подрезки 2 как раз хорошо центрируется вокруг атома кислорода воды, но не вокруг ближайшего углерода смещенного гистидина. ЭП же вокруг азота NE2 в изначальном положении остатка совпадает с его центром, но радиус отображения ЭП меньше - уровень в 2 сигмы достигается очень близко к центру.

Я предполагаю два возможных объяснения наблюдаемой картине. Во-первых, вода могла действительно присутствовать в кристалле на этом месте, а PDB Redo удаляет ее и вписывает на ее место имидазольное кольцо, но центры его атомов не совпадают с локальным максимумом ЭП. Во-вторых, независимо от наличия воды в данном месте в какой-то доле ячеек, оба положения гистидина могут часто встречаться в кристалле и представлять собой альтернативные конформации. В пользу этого объяснения говорит также то, что ЭП на более низких уровнях подрезки “повторяет” форму имидазольного кольца. Не могу сказать, улучшил ли PDB Redo отображение данного остатка или нет - ни с точки зрения наложения на ЭП, ни с позиции биологического смысла.

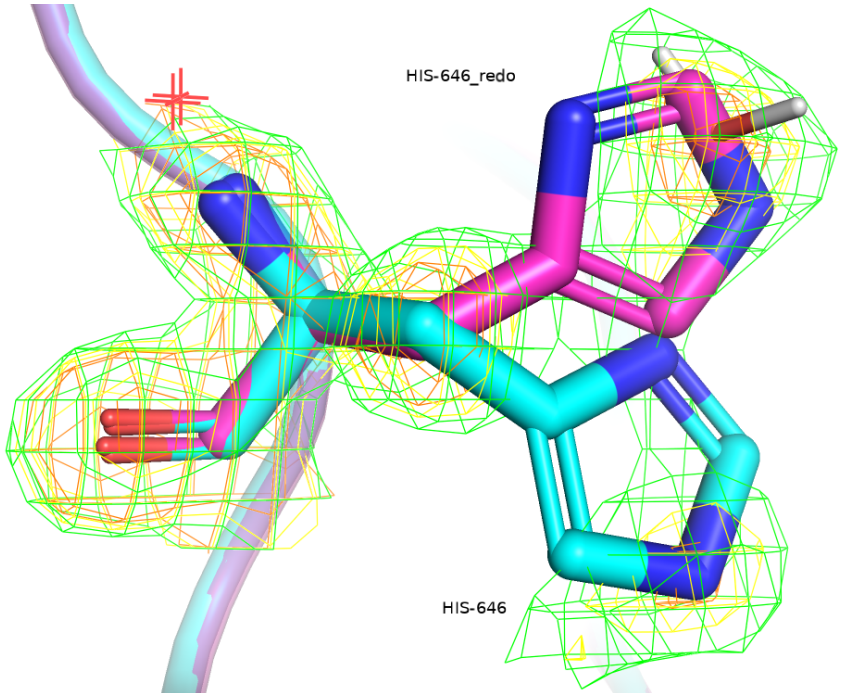

Рисунок 17. Гистидин-646 исходной структуры (синий), гистидин-646 измененной структуры (фиолетовый). ЭП зеленым - 1 сигма, желтым - 1.5, оранжевым - 2.

4.2. Отчеты валидации

Качество модели после изменений сервисом PDB Redo существенно улучшилось. R и R_{free} уменьшились до 0.124 и 0.151 соответственно. Число выбросов на карте Рамачандрана уменьшилось до 2, было изменены типичные углы связи для 109 остатков, для 305 остатков показано улучшение их соответствия электронной плотности, для 2 остатков - ухудшение (Рисунок 19). Z-score для карты Рамачандрана, ротамеров и R_{free} улучшились по сравнению со средними значениями моделей со схожими разрешениями (Рисунок 20).

Я также загрузил исходную структуру в сервис MolProbity и привел сводную таблицу (Рисунок 21) по всем атомам после добавления водородов (без внесения инверсий остатков). Видно, что по сравнению с изначальной моделью (Рисунок 12) улучшения получены по большинству параметров. На порядок сократилось количество остатков, маргинальных по длине связи или углу связи. Как отмечено и в отчете PDB Redo Z-score карты Рамачандрана сильно снизился и находится около 0. Значительно снизилось количество наложений, clashscore упал более чем в два раза до 3.2 единиц и находится в числе 97% лучших среди структур с похожим разрешением. Отчет WHAT_CHECK сообщает, что число клэшей сократилось с 575 до 141. При этом для измененной структуры указывается в 2 раза больше ошибочных положений ASN, GLN, HIS. В этом же отчете указаны т.н. “troublesome residues”, которые заслуживают визуальной инспекции. Не знаю, насколько эта метрика достоверна, но после изменения PDB Redo число таких остатков снизилось с 1145 до 327.

Далее я просмотрел отчет MolProbity после изменения модели в PDB Redo на предмет маргинальности остатков с аннотированной функцией и сравнил с исходной моделью без каких-либо преобразований (Таблица 2). Видно, что исправление модели уточнило параметры связи и в функциональных остатках такие отклонения ушли, однако, некоторые все равно накладываются на соседние атомы, но такие наложения минимальны и описаны лишь для некоторых цепей.

Стоит сказать и о ионах металлов в данной модели. Выдачи CheckMyMetal до и после изменения в PDB Redo см. на странице практикума. Согласно отчетам, после изменения структуры количество аутлаеров среди ионов металлов существенно сократилось, в основном по параметру VECSUM [3], который отражает симметричность координационной сферы иона, и параметру Valence, что для катиона металла аналогично заряду [4]. Рассмотрим сдвиг положения металла на примере атома магния 3105

(заселенность 0.5) из цепи A (Рисунок 18), который координирован 5 молекулами воды и кислородом остова аспарагина-597. Видно, что атом в модели после изменения занимает более центральное положение по отношению к координирующим атомам (в центре октаэдра). В отчете CMM это отразилось в улучшении метрики VECSUM. Также после изменения модели CMM предлагает альтернативный металл на эту позицию - кобальт или медь.

Подводя итог, предположу, что измененную модель можно использовать для изучения особенностей белка с большей уверенностью, так как качество возросло, количество маргиналов существенно упало, остатки с аннотированной функцией перестали быть маргинальными, а ухудшения незначительны. При этом необходимо рассматривать изучаемые области и остатки индивидуально, ведь нам необходимо работать с реалистичной, осмысленной, моделью, отражающей биологические функции, а не с моделью, которая лучше подстроена под экспериментальные данные.

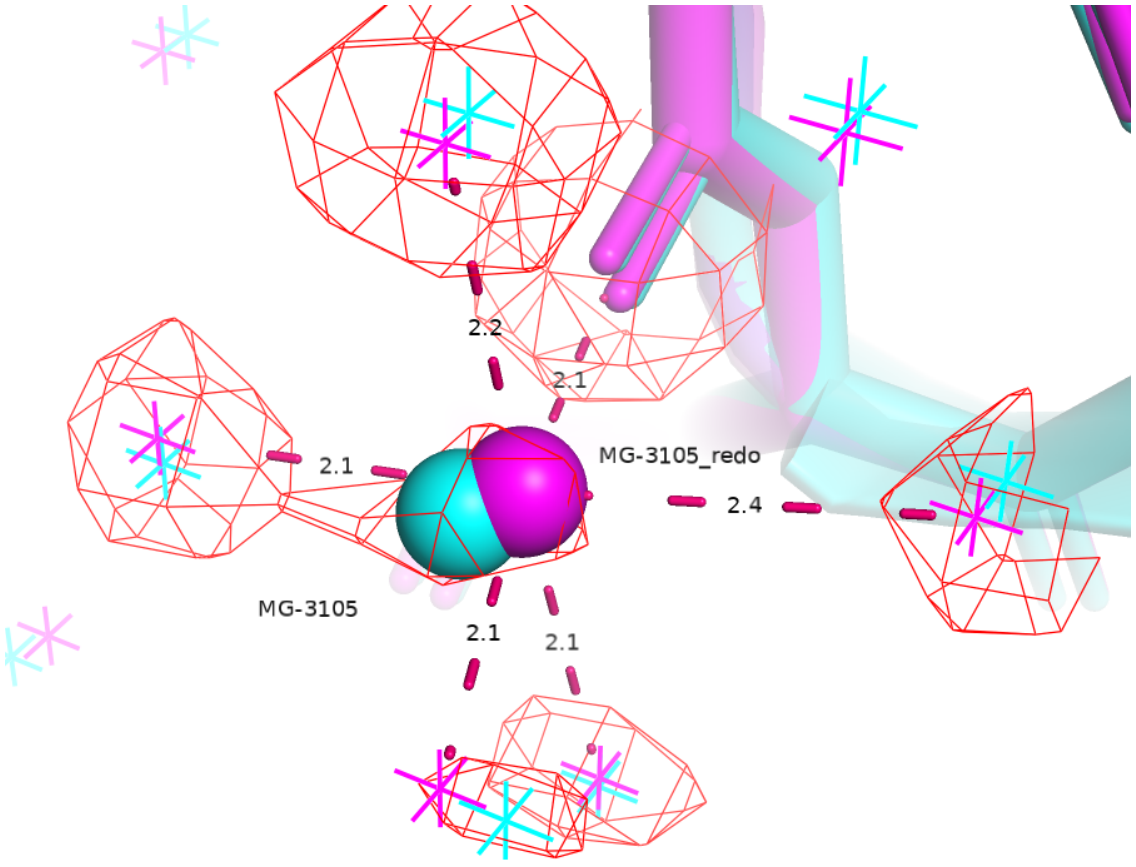

Рисунок 18. Ион магния до исправления (голубой) и после (фиолетовый) со своим окружением. Крестиками показаны молекулы воды.

Significant model changes

Description	Count
Rotamers changed	109
Side chains flipped	0
Waters removed	168
Peptides flipped	1
Chiralities fixed	0
Residues fitting density better	305
Residues fitting density worse	2

Рисунок 19. Основные улучшения. внесенные PDB Redo.

Model quality compared to resolution neighbours

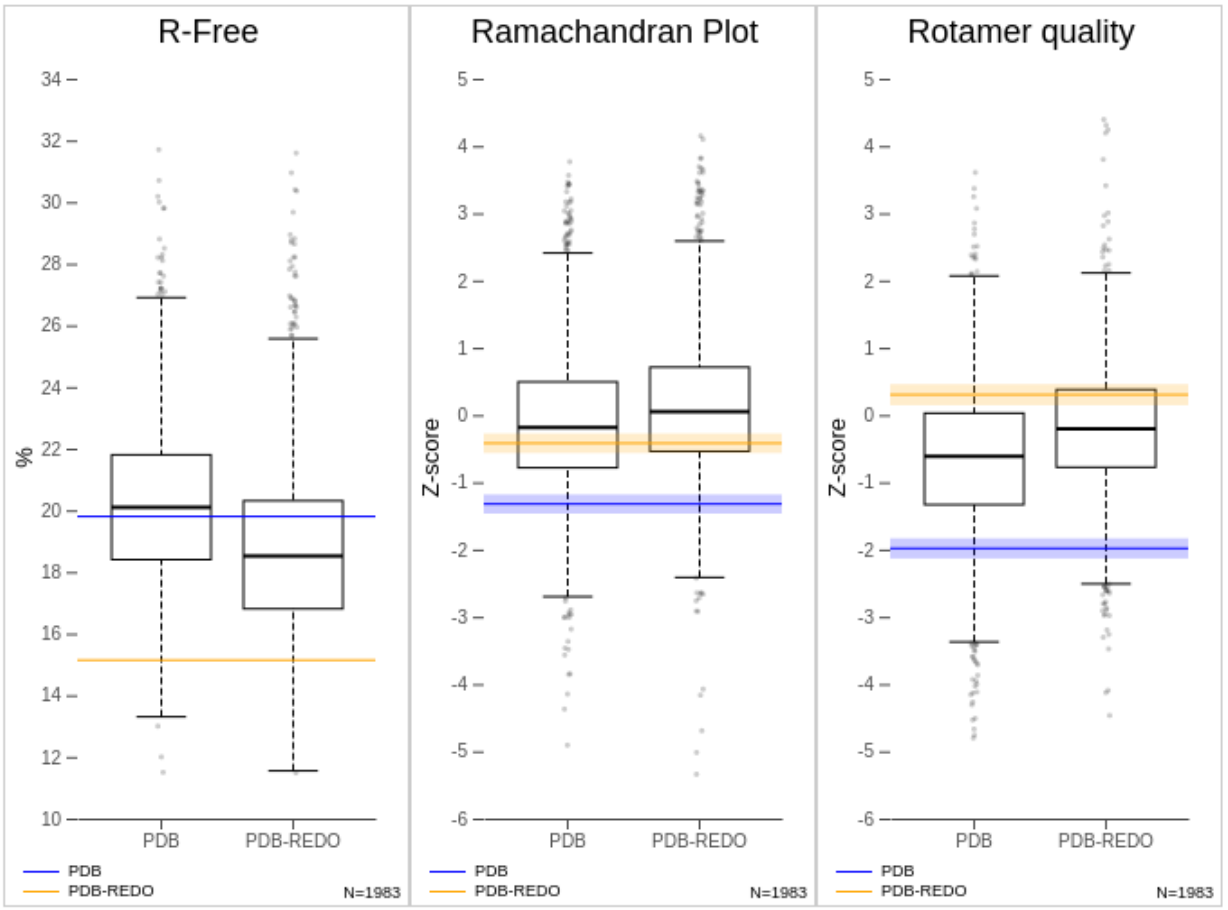

Рисунок 20. Качество измененной модели по сравнению с исходной.

All-Atom Contacts	Clashscore, all atoms:	3.22		97th percentile* (N=598, 1.50Å ± 0.25Å)
	Clashscore is the number of serious steric overlaps (> 0.4 Å) per 1000 atoms.			
Protein Geometry	Poor rotamers	30	0.87%	Goal: <0.3%
	Favored rotamers	3332	96.52%	Goal: >98%
	Ramachandran outliers	2	0.05%	Goal: <0.05%
	Ramachandran favored	3924	97.51%	Goal: >98%
	Rama distribution Z-score	-0.17 ± 0.13		Goal: abs(Z score) < 2
	MolProbity score^	1.21		97th percentile* (N=4836, 1.50Å ± 0.25Å)
	Cβ deviations >0.25Å	11	0.29%	Goal: 0
	Bad bonds:	41 / 33849	0.12%	Goal: 0%
	Bad angles:	52 / 46085	0.11%	Goal: <0.1%
Peptide Omegas	Cis Prolines:	20 / 248	8.06%	Expected: ≤1 per chain, or ≤5%
	Cis nonProlines:	12 / 3792	0.32%	Goal: <0.05%
Additional validations	Tetrahedral geometry outliers	2		
	Waters with clashes	222/4046	5.49%	See UnDowser table for details

Рисунок 21. Сводная таблица MolProbity по измененной структуре.

Таблица 2. Остатки с аннотированной функцией и их качество.

Ам. к-та	№ в UniProt/PDB	Аннотированная функция	Отчет MolProbity по исходной структуре (без инверсий)	Отчет MolProbity по измененной структуре
N	103/102	Связывание субстрата	Клэш	Клэш
D	202/201	Связывание субстрата	Маргинал по углам, длине связи	-
H	358/357	Стабилизация ПС	Клэш, маргинал по углу, длине связи	-
H	392/391	Стабилизация ПС	Маргинал по длине связи	Клэш
E	417/416	Связывание магния	Маргинал по углам, клэш	-
H	419/418	Связывание магния	Клэш, маргинал по длине связи и углам	-
E	462/461	Донор протона в активном центре	Маргинал по длине связи и углам	-
E	538/537	Нуклеофил в активном центре	2 клэша, маргинал по длине связи	Клэш
N	598/597	Связывание магния	Маргинал по углам	Клэш
F	602/601	Связывание натрия	Клэш, маргинал по углам	Клэш

N	605/604	Связывание субстрата	Минимальный клэш	-
W	1000/999	Регуляция	Клэш	-

Источники

1. Bank RPD. 1JZ7. [cited 1 Dec 2021]. Available: https://www.rcsb.org/structure/1jz7

2. Juers DH, Heightman TD, Vasella A, McCarter JD, Mackenzie L, Withers SG, et al. A structural view of the action of Escherichia coli (lacZ) beta-galactosidase. Biochemistry. 2001;40: 14781–14794.

3. Müller P, Köpke S, Sheldrick GM. Is the bond-valence method able to identify metal atoms in protein structures? Acta Crystallogr D Biol Crystallogr. 2003;59: 32–37.

4. Brown ID. Recent developments in the methods and applications of the bond valence model. Chem Rev. 2009;109: 6858–6919.