**MD-dataset课题进展汇报**

周兆寅

日期：2025.7.14 ~ 2025.7.19

**一、工作重点**

1. 标注D3PM数据库，挑选简单体系优先进行MD模拟。

**二、课题背景与思路**

目前缺少标准的蛋白-配体复合物体系长时间MD模拟数据集，现有数据集更关注蛋白质本身的结构运动，而非药物设计中关心的配体与蛋白的结合相互作用。

彭诚师兄的D3PM数据库1包含大量同一个蛋白在配体结合前后的构象对（一个apo与一个complex为一对），其中865对的口袋残基构象变化显著。本课题计划基于D3PM库中收集的晶体结构，优先选取简单体系进行模拟，获得空蛋白以及蛋白-配体复合物体系长时间的MD轨迹数据集。

**三、实验目的**

* 模拟apo-protein口袋的结构系综，探究晶体结构与溶液中优势构象的差异；
* 比较apo-protein的结构系综与complex晶体结构中的配体诱导契合口袋构象；
* 为训练AI模型进行数据储备。

**四、实验结果**

1. 标注D3PM数据库

目前缺少标准的蛋白-配体复合物体系长时间MD模拟数据集，现有数据集更关注蛋白质本身的结构运动，而非药物设计中关心的配体与蛋白的结合相互作用。

1. 体系准备（MD自动化流程的输入文件）

* 蛋白，配体质子化条件：依据晶体结构解析的pH而定，若未提及，使用生理条件（pH=7.4）
* 蛋白补链程序：Amber pdb4amber；pdb2pqr；Schrödinger protein preparation
* 配体质子化程序：Schrödinger ligprep Epik
* 力场未覆盖的离子（如硫酸根离子）参数准备：

根据口袋环境确定质子化状态，高斯（g16）优化后拟合resp电荷，Ambertools antechamber生成prep与frcmod文件

1. MD体系构建（MD自动化流程自动进行）

* 配体高斯优化关键词：B3LYP/6-31G\* em=gd3bj pop=MK iop(6/33=2,6/42=6) opt
* 力场参数：Amber ff14SB, TIP3P, GAFF2
* 体系溶剂盒子大小：向外延伸12Å
* 预平衡中的位置限制：

能量最小化不进行位置限制；nvt，npt平衡中蛋白，配体全原子均施加位置限制

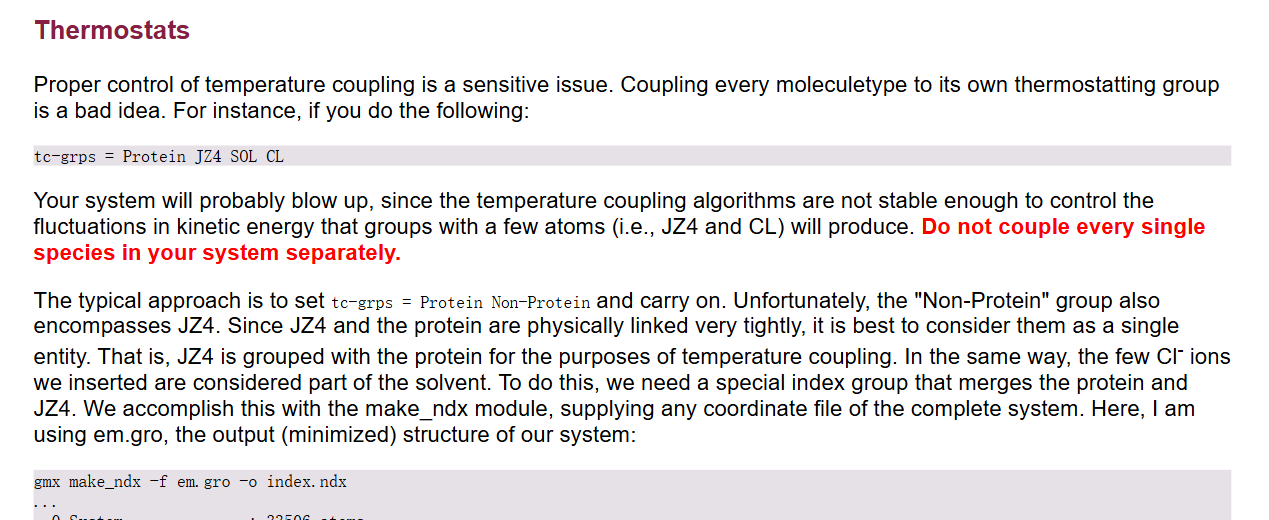
* CMD模拟软件：GMX 2024.4
* 预平衡中的部分关键参数（mdp）

1. 正则系综（nvt）平衡

* **温度：310K？300K？**

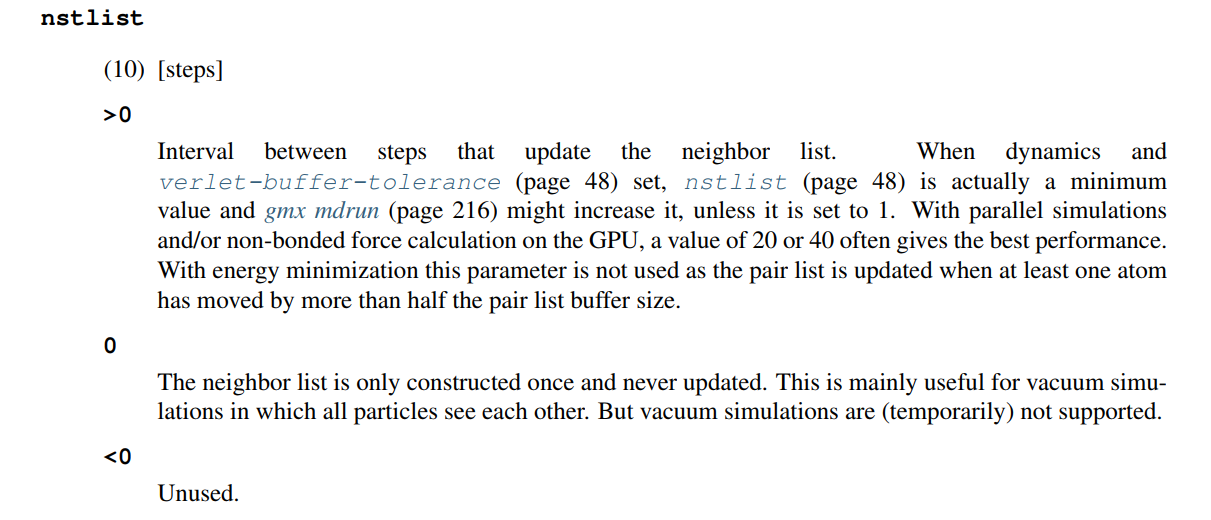
该参数需要讨论，王进安老师建议使用人体常温（37℃，310K）。TIP3P力场的参考温度尚未找到文献依据。

* tc-grps（温控组）设置为溶质与溶剂两个组，更加精确



* nstlist（更新邻区列表的频率）= 20 （步）

手册提到当使用GPU进行并行计算和(或)非键作用计算时, 该值取20或40往往能得到最佳性能



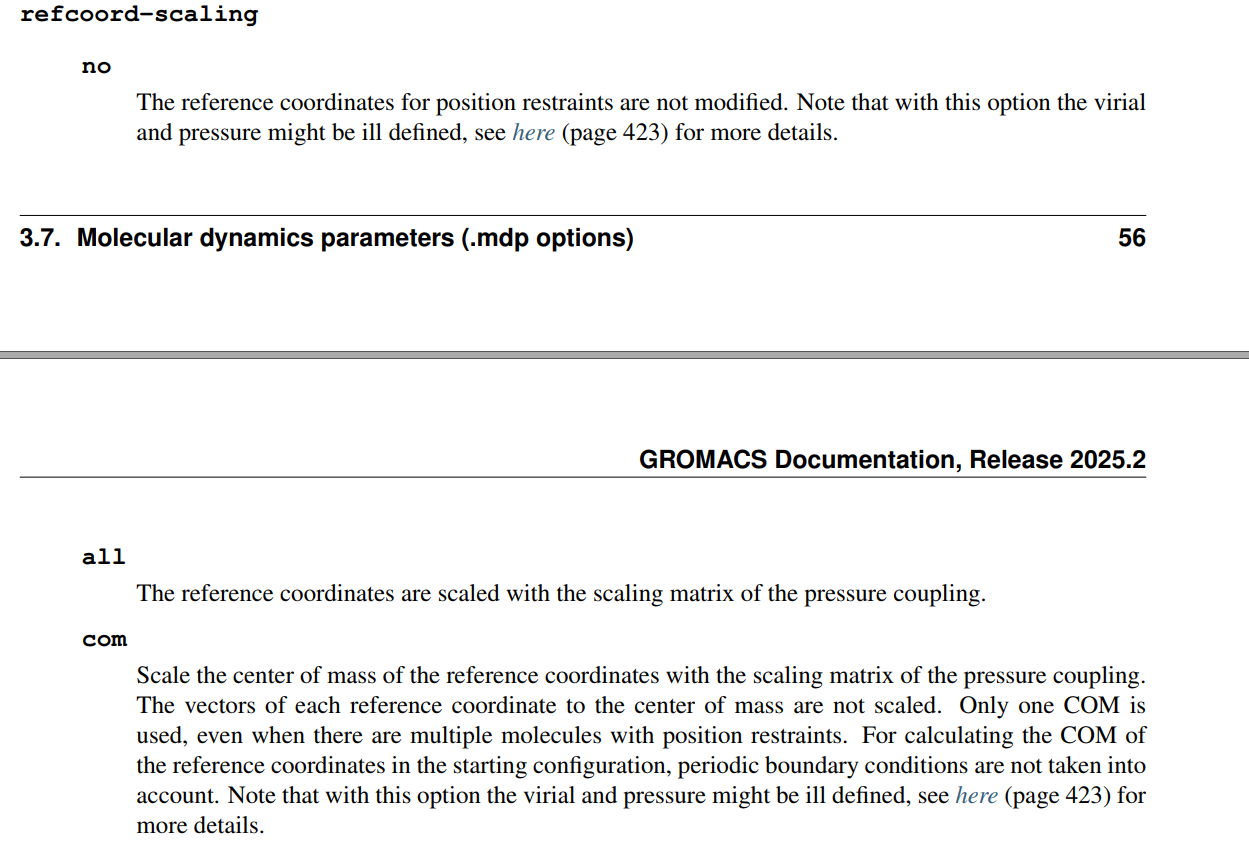
1. 等温等压系综（npt）预平衡

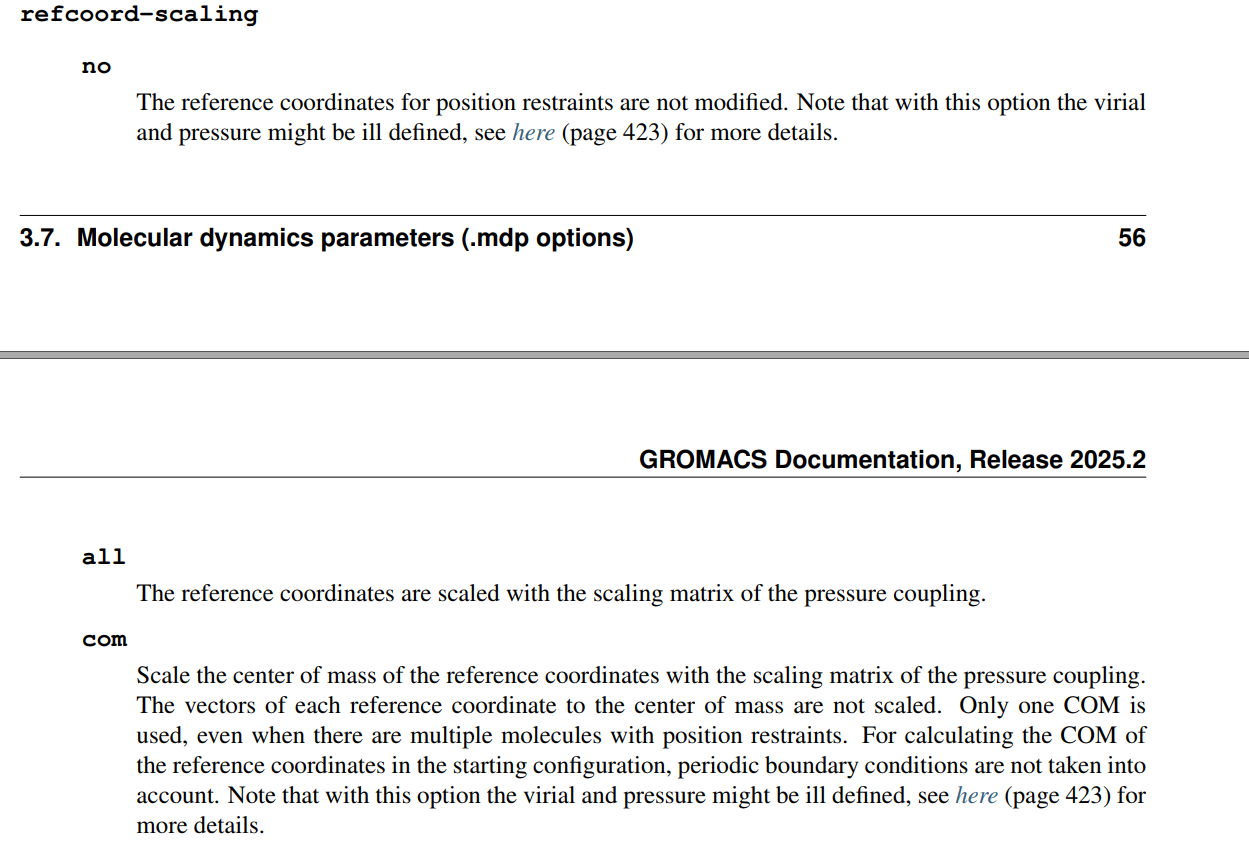
* **温度：310K？300K？**

该参数需要讨论，王进安老师建议使用人体常温（37℃，310K）。TIP3P力场参数的参考温度尚未找到文献依据。

* refcoord\_scaling = com

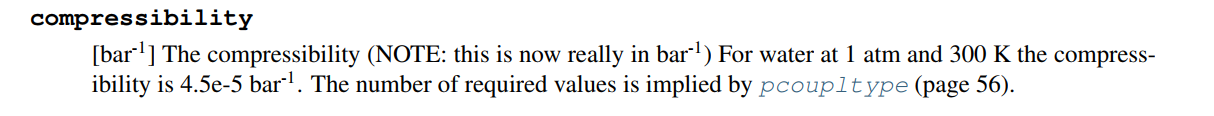
在预平衡时，我们往往对某些原子进行位置限制（常用的是蛋白+配体）。这个参数下，我们首先在不考虑周期性条件的情况下计算进行了位置限制所有原子的质心。控压算法调节盒子尺寸时，按照盒子比例缩放这个质心（我的理解是直接缩放三维坐标），同时保证进行了位置限制的每个原子与该质心的矢量不变（距离，相对位置等），该参数相当于依照质心对施加位置限制的原子参考坐标进行统一的空间位移。





* nstlist（更新邻区列表的频率）= 20 （步）与nvt一致
* compressibility （压缩系数）= 4.5e-5？4.4e-5

如图1所示，压缩系数与溶剂类型及温度密切相关。1个标准大气压, 300 K条件下的水压缩系数为4.5e–5 [bar-1]；1个标准大气压, 310 K条件下的水压缩系数为4.4e–5 [bar-1]。不同溶解之间的压缩系数差别极大。后续该参数的值可参考文献Kell, George S. "Isothermal compressibility of liquid water at 1 atm." *Journal of Chemical and Engineering Data* 15.1 (1970): 119-122.进行设置2。



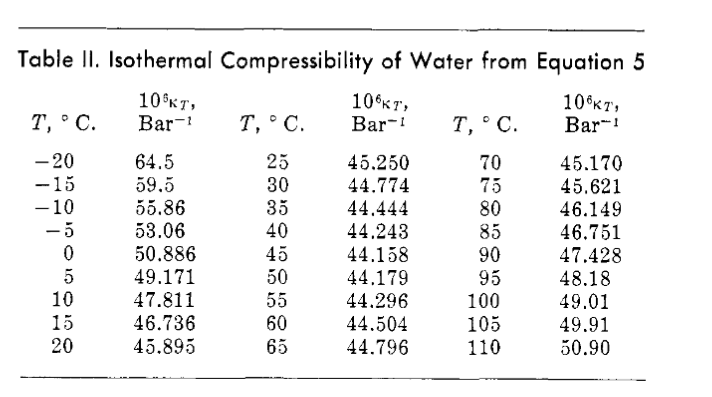


图1. 水在不同温度下的压缩系数2

* 1μs CMD模拟的部分关键参数（mdp）
* nstxout / nstvout / nstfout = 1000（步）

输出坐标，速度与受力信息到trr文件。该轨迹文件相较于xtc文件体积更大，但包含了速度，受力这些xtc文件没有的信息，可用于训练AI模型。

* gen\_vel = yes & gen\_seed = -1

按照Maxwell分布产生随机初速度，设置随机种子。便于对同一体系进行三次平行模拟。

* compressibility，nstlist等参数与预平衡保持一致，在此不再赘述

**四、总结&后续计划**

1. MD数据集将通过人工挑选模拟体系并准备输入文件-MD自动化脚本进行MD体系构建-提交CMD任务的流程逐步进行构建。通过与乐云师姐的讨论，基本确定了模拟参数；
2. **模拟参数中的温度设置为300K还是310K需要讨论**；
3. MD-automation系列脚本已可以实现体系的自动化准备，主要的耗时环节为体系的挑选与输入文件准备。待模拟参数敲定后，计划每周准备 1-4个体系（根据空闲的计算资源决定）。

1. Peng, C.; Zhang, X.; Xu, Z.; Chen, Z.; Yang, Y.; Cai, T.; Zhu, W., D3PM: a comprehensive database for protein motions ranging from residue to domain. *BMC bioinformatics* **2022,** *23* (1), 70.

2. Kell, G. S. J. J. o. C.; Data, E., Isothermal compressibility of liquid water at 1 atm. **1970,** *15* (1), 119-122.

参考文献：