PATTER, Volume *1*

Supplemental Information

A Searchable Database of Crystallization Cocktails in the PDB: Analyzing the Chemical Condition Space

Miranda L. Lynch, Max F. Dudek, and Sarah E.J. Bowman

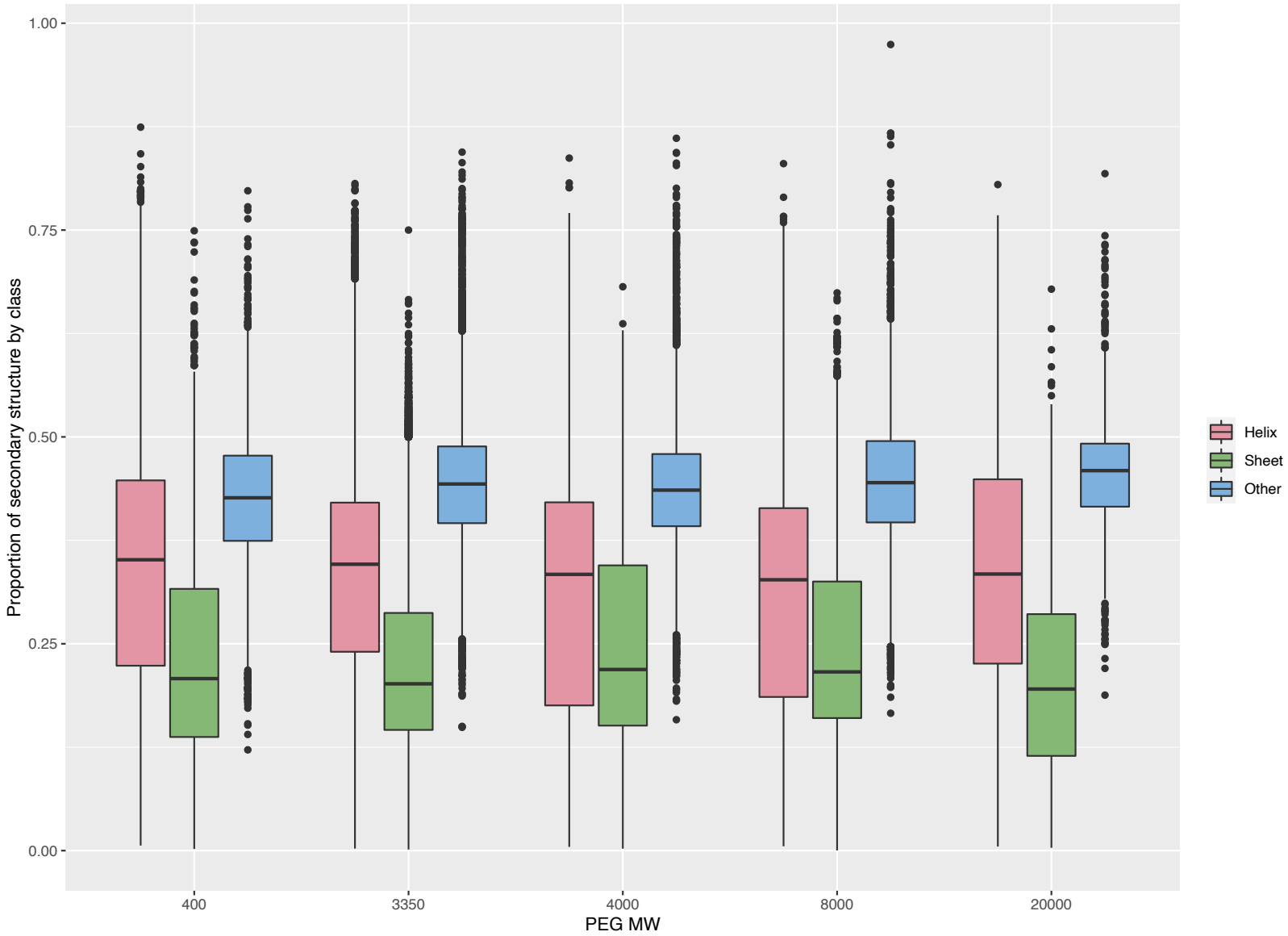

Figure S1: Boxplots of proportions of secondary structural elements by class, for a subset of PEG MW. The PEG MWs selected (PEG 400, 3350, 4000, 8000, and 20000) were chosen as these are common PEG cocktail components, and nearly span the full range of possible PEG values. Each PEG MW is associated with a trio of boxplots, one for each secondary structure element (which must sum to one for a given PDB ID). Each PEG MW is associated with a unique number of PDB IDs that used that PEG in the crystallization cocktail. Even in these box plots of a subset of PEG MW, the trends observed in the full model are visible, with decreasing *helix* content and increasing *other* proportion of secondary structure associated with increasing PEG MW.

Crystallization Conditions	Details Extracted from PDB	Details Parsed	CDD
REMARK 280 CRYSTALLIZATION CONDITIONS: PEG 3350, PH 8, VAPOR DIFFUSION, REMARK 280 HANGING DROP, TEMPERATURE 290K	PEG 3350, pH 8, VAPOR DIFFUSION, HANGING DROP, temperature 290K	['PEG 3350', None]	✓
REMARK 280 CRYSTALLIZATION CONDITIONS: 15 % W/V POLYETHYLENE GLYCOL PEG400, REMARK 280 30 % W/V PEG 1500 AND 0.1 M HEPES, PH 7.5, VAPOR DIFFUSION, REMARK 280 SITTING DROP, TEMPERATURE 289K	15 % w/v polyethylene glycol PEG400, 30 % w/v PEG 1500 and 0.1 M Hepes, pH 7.5, VAPOR DIFFUSION, SITTING DROP, temperature 289K	['PEG 400', '15% w/v', 'PEG 1500', '30% w/v', 'HEPES', '100.0']	✓
REMARK 280 CRYSTALLIZATION CONDITIONS: CONTAINERLESS BATCH METHOD: 300 REMARK 280 MICROL OF HIGH-DENSITY SILICON OIL WAS TRANSFERRED INTO A REMARK 280 WELL OF A CELL CULTURE PLATE AND OVER-LAID WITH 500 MICROL REMARK 280 OF REGULAR SILICON OIL. AT THE INTERFACE BETWEEN THE TWO REMARK 280 LIQUIDS A DROPLET WAS DEPOSITED THAT WAS OBTAINED BY REMARK 280 MIXING 0.4 MICROL OF H2O, 2.2 MICROL OF 5BETA-POR SOLUTION REMARK 280 WITH NADP AND PROGESTERONE ADDED AND 1.8 MICROL OF A REMARK 280 CRYSTALLIZATION SOLUTION CONSISTING OF 22.9 % MPD, 3.5 % REMARK 280 PEG 8000, 0.05 M SODIUM ACETATE, 0.02 M CACL2, PH 5.8.	CONTAINERLESS BATCH METHOD: 300 MICROL OF HIGH-DENSITY SILICON OIL WAS TRANSFERRED INTO A WELL OF A CELL CULTURE PLATE AND OVER-LAID WITH 500 MICROL OF REGULAR SILICON OIL. AT THE INTERFACE BETWEEN THE TWO LIQUIDS A DROPLET WAS DEPOSITED THAT WAS OBTAINED BY MIXING 0.4 MICROL OF H2O, 2.2 MICROL OF 5BETA-POR SOLUTION WITH NADP AND PROGESTERONE ADDED AND 1.8 MICROL OF A CRYSTALLIZATION SOLUTION CONSISTING OF 22.9 % MPD, 3.5 % PEG 8000, 0.05 M SODIUM ACETATE, 0.02 M CACL2, PH 5.8.	['MPD', '22.9%', 'PEG 8000', '3.5%', 'sodium acetate', '50.0', 'calcium chloride', '20.0']	✓
REMARK 280 CRYSTALLIZATION CONDITIONS: AMMONIUM SULPHATE, POTTASIUM REMARK 280 PHOSPHATE, POTASSIUM CHLORIDE, PH 6.8, VAPOR DIFFUSION, HANGING REMARK 280 DROP, TEMPERATURE 277K	AMMONIUM SULPHATE, POTTASIUM PHOSPHATE, POTASSIUM CHLORIDE, pH 6.8, VAPOR DIFFUSION, HANGING DROP, temperature 277K	['ammonium sulfate', None, 'potassium phosphate', None, 'potassium chloride', None]	✓
_exptl_crystal_grow.pdbx_details '100 mM HEPES pH 7.5 and 80% 2-Methyl-2,4-pentanediol (MPD)'	100 mM HEPES pH 7.5 and 80% 2-Methyl-2,4-pentanediol (MPD)	['HEPES', '100.0', 'MPD', '80%']	✓

Table S1: Examples of successful parsing from the PDB Crystallization Conditions free text field to the CDD. Data is extracted from the crystallization details free text field and the components are parsed using controlled vocabularies to generate the CDD. The detail parsing function extracts chemical names, even when concentration information is not present, as exemplified in the top and 4th row. Chemical concentrations are extracted when present and converted to mM units or specified as %v/v or %w/v. The detail parsing function is able to parse significant text present in the free text field, as shown in the 3rd row, and can fix spelling errors as shown in the 4th row. We note that the data extraction works with both the REMARK 280 (PDB format) and the exptl_crystal_grow.pdbx_details (mmCIF file format) fields, as shown in the bottom row. These five PDB IDs are included in the 99,229 PDB IDs in the CDD.

Crystallization Conditions	Details Extracted from PDB	Details Parsed	Parsing Issue
REMARK 280 CRYSTALLIZATION CONDITIONS: NACL, MPD, PEG, PH 5.5, VAPOR REMARK 280 DIFFUSION, HANGING DROP, TEMPERATURE 298K	NaCl, MPD, PEG, pH 5.5, VAPOR DIFFUSION, HANGING DROP, temperature 298K	['sodium chloride', None, 'MPD', None, 'PEG ', None]	Dictionary does not recognize the word 'PEG' by itself
REMARK 280 CRYSTALLIZATION CONDITIONS: 21-35% PEG 3350, 0.1 M BIS-TRIS PH 5.5 REMARK 280 -7.0 OR 21-40% MEDIUM-MOLECULAR-WEIGHT PEG SMEARS (MMW PEG REMARK 280 SMEARS) BUFFERED EITHER WITH 0.1 M BIS-TRIS PH 6.0-7.5 OR 0.1 M REMARK 280 TRIS PH 7.5-8.8, VAPOR DIFFUSION, SITTING DROP, TEMPERATURE REMARK 280 277.15K	21-35% PEG 3350, 0.1 M Bis-Tris pH 5.5-7.0 or 21-40% medium-molecular-weight PEG smears (MMW PEG smears) buffered either with 0.1 M Bis-Tris pH 6.0-7.5 or 0.1 M Tris pH 7.5-8.8	['PEG 3350', '28.0%', 'medium-molecular-weight', '30.5%', 'PEG ', '6.25', 'smears mmw PEG smears', None, 'bis-tris', '100.0', 'tris', '100.0']	Details provided are not specific enough. The dictionary does not recognize the word 'or', nor does it recognize 'smears'
REMARK 280 CRYSTALLIZATION CONDITIONS: NULL, VAPOR DIFFUSION, SITTING DROP, REMARK 280 TEMPERATURE 293K	Null	['null', None]	No details provided
REMARK 280 CRYSTALLIZATION CONDITIONS: THE PURIFIED PROTEIN WAS USED IN REMARK 280 CRYSTALLISATION TRIALS EMPLOYING BOTH, A STANDARD SCREEN WITH REMARK 280 APPROXIMATELY 1200 DIFFERENT CONDITIONS, AS WELL AS REMARK 280 CRYSTALLISATION CONDITIONS IDENTIFIED USING LITERATURE DATA., REMARK 280 BATCH MODE, TEMPERATURE 293K	The purified protein was used in crystallisation trials employing both, a standard screen with approximately 1200 different conditions, as well as crystallisation conditions identified using literature data.	['trials employing', None, 'standard', None, 'different', '1200', 'identified', None, 'literature', None]	No details are provided, and the dictionary does not recognize any of the words as chemical components

Table S2: Examples of unsuccessful parsing from the PDB Crystallization Conditions free text field. The detail parsing function is unable to parse the details listed in the free text field in 34,508 PDB IDs, and these four PDB IDs provide examples of some of the myriad reasons why the details are not able to be automatically parsed.