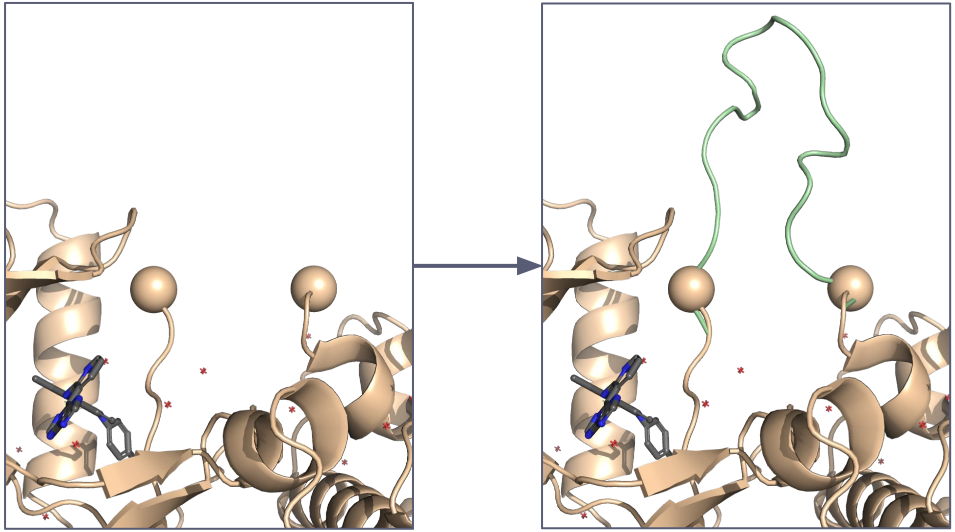
Modelling of missing loops in protein structures

Jan-Oliver Kapp-Joswig

May 2025

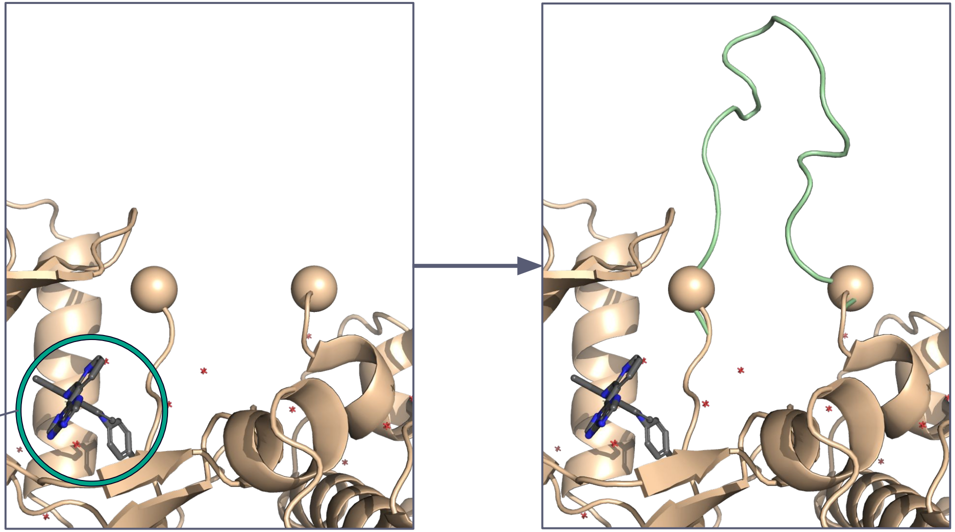
Problem statement

- Input: partial structure + sequence
- Output: structure(s) with filled gaps
- Requirements:
  - physical plausibility
  - conformational variability
  - confidence score



Problem statement

- Input: partial structure + sequence
- Output: structure(s) with filled gaps
- Requirements:
  - physical plausibility
  - conformational variability
  - confidence score



Relevant example:
Binding site close to missing loop

General modelling strategies

- **Template-based/homology modelling:**
  - ProMod3 (OpenStructure), Modeller
- **Fragment-based (using a database):**
  - FREAD
- **Ab initio modelling:**
  - PDBFixer, ICM (Molsoft), Rosetta, Modeller
- **ML models:**
  - AF2 (“inpainting” mode)
- **Hybrid approaches**

General modelling strategies

- Template-based/homology modelling:
  - ProMod3 (OpenStructure), Modeller

- Relies on (multiple) sequence and structural alignment
- Good if high quality references exist
- In principle, very long segments can be built
  (but in practice finding good templates becomes difficult)

General modelling strategies

- **Fragment-based (using a database):**
  - FREAD

- **Assumes that similar sequences adopt similar conformations**
- **Good if suitable fragments exist (database completeness)**
- **Can struggle with longer segments**

General modelling strategies

- **Ab initio modelling:**
  - PDBFixer, ICM (Molsoft), Rosetta, Modeller

- Modelling without prior knowledge from physical principles
- Several algorithms available: CCD, KIC, MC chain growth
- Can struggle with longer segments

General modelling strategies

- **ML models:**
  - **AF2 (“inpainting” mode)**

- **Guided predictions on partially solved input can have improved outcomes**
- **Quality of results are system dependent**

In practice

- Leverage multiple available tools
  - Compare, filter, and prioritise models

- Generate model ensembles
  - Averaging
  - Clustering

- Refine models (Energy minimisation, MD)

- Benchmark approaches on problems with known solution

Evaluation of results

- Some tools provide a built-in score:
    - AF2

- Geometric scores (discard unphysical results):
    - clashes
    - bond length/angle deviations from expected values
    - Ramachandran outliers

- Energy functions (rank quality among plausible results):
    - Statistical potentials
    - Force fields
    - Solvation energy

Challenges

- Accessibility of tools can be a hindrance
    - included in large suites
    - GUI based
    - web servers
    - missing documentation
    - licences
    - dependencies

- PDBx/mmCIF is a complex format
    - parsing, modification, splitting/joining
    - tools like `gemmi` help

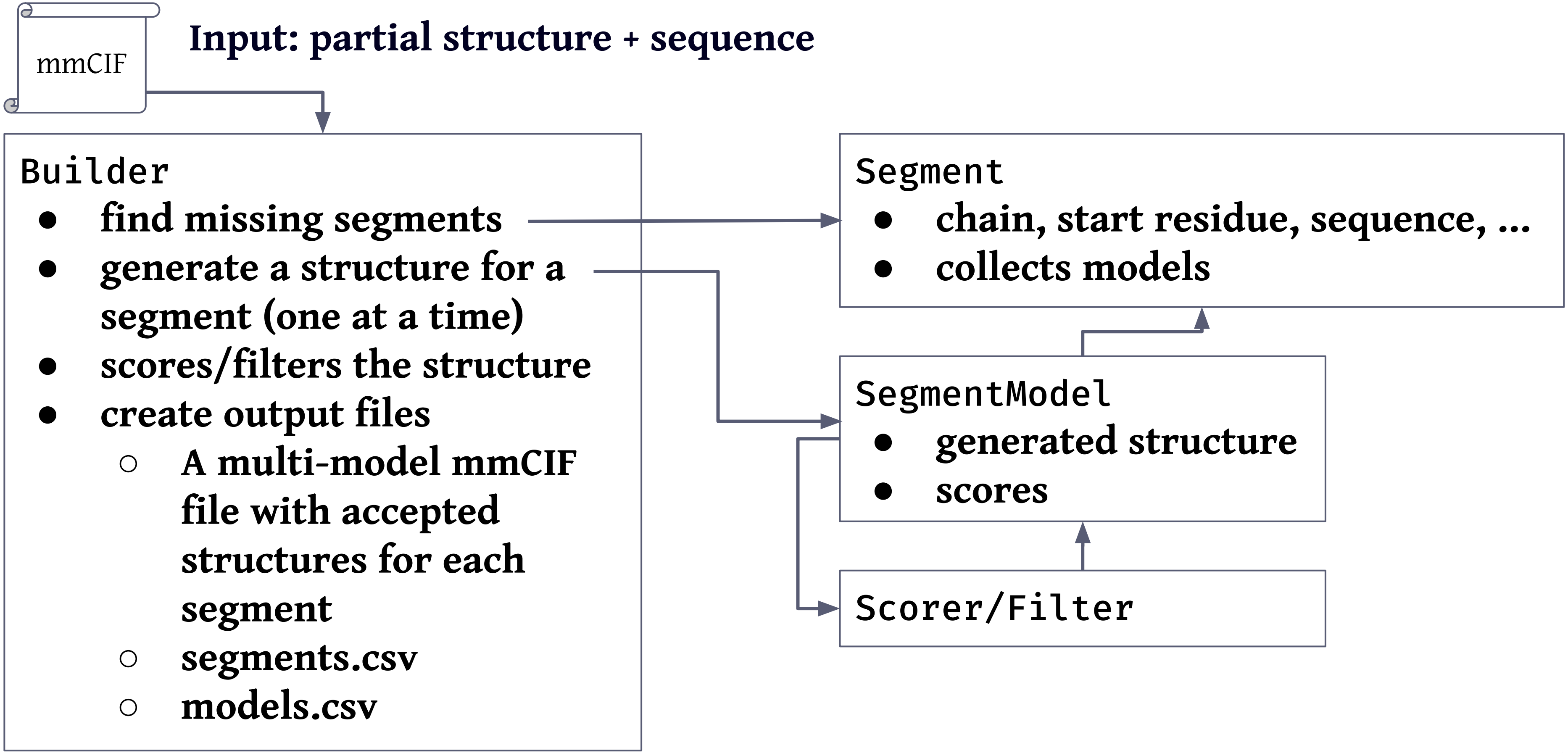
Minimal approach

- `LoopBuilder` Python package
- Lightweight skeleton to generalise the use of external tools
  - to generate structures for missing segments
  - to score individual segments

- Pick simple methods as baseline
  - PDBFixer
  - MolProbity

- Opportunity to
  - establish general protocol
  - sort technical details
  - identify bottlenecks

Minimal approach



mmCIF

Input: partial structure + sequence

`Builder`

- find missing segments
- generate a structure for a segment (one at a time)
- scores/filters the structure
- create output files
  - A multi-model mmCIF file with accepted structures for each segment
  - segments.csv
  - models.csv

`Segment`

- chain, start residue, sequence, ...
- collects models

`SegmentModel`

- generated structure
- scores

`Scorer/Filter`

Example: 3IDP

```
builder = PDBFixerBuilder(
    structure_file=PROJECT_ROOT / "data/3idp.cif",
    output_directory=PROJECT_ROOT / "sandbox/PDBFixer",
    scorers=[MolProbityScorer(docker_image="francecosta/molprobity:v0.0.1")],
    filters=[lambda x: x.scores.get("ramachandran_outliers", 0.3) <= 0.3],
    working_directory=PROJECT_ROOT / "sandbox/PDBFixer/tmp",
)
builder.build(n=3)
```

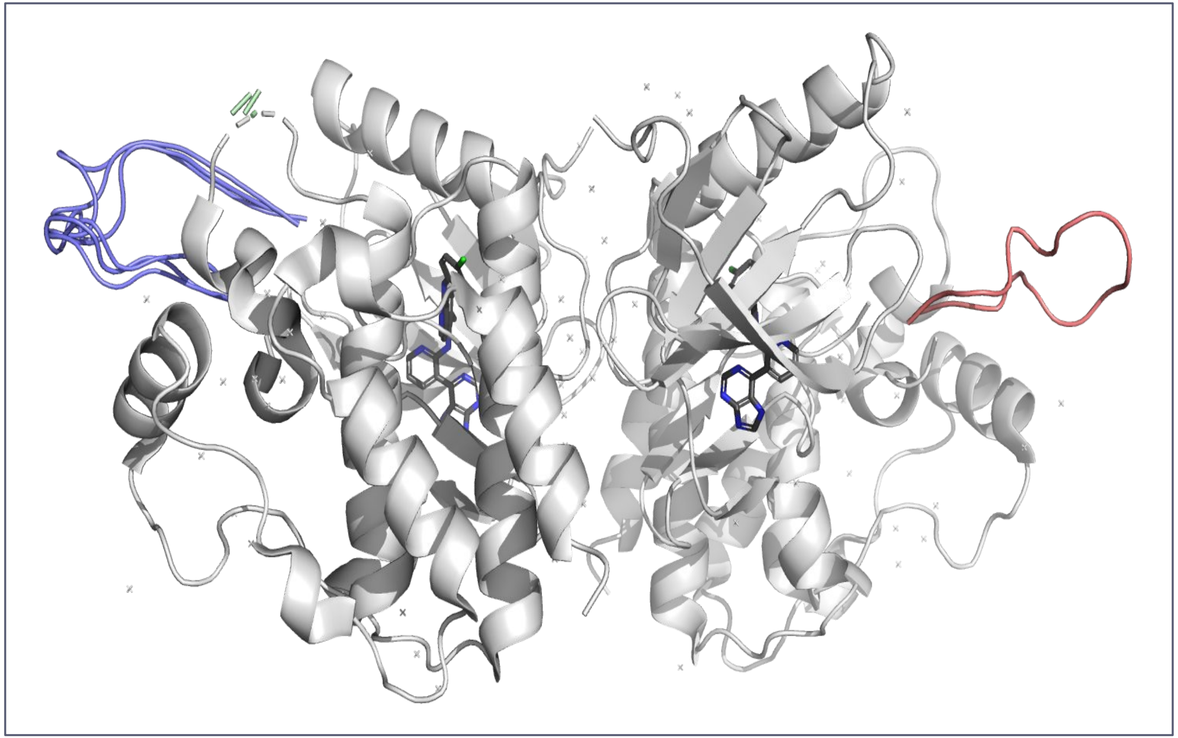
- **segments.csv**

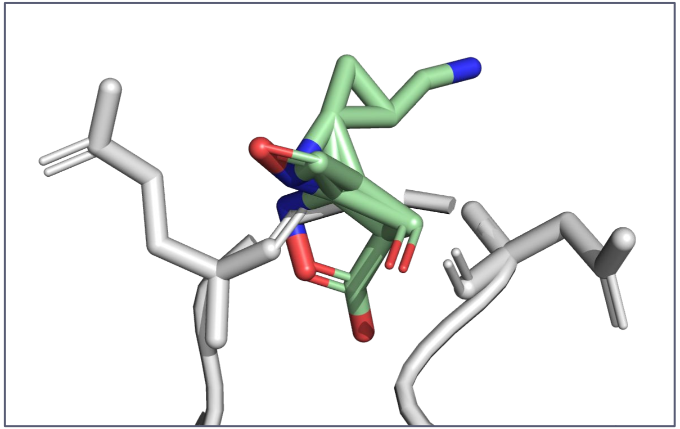
```
identifier,chain_index,chain_name,residue_start_index,residue_start_seqid,residue_index_offset,residue_names, ...
loop_1,0,A,149,598,449,"['ALA', 'THR', 'GLU', 'LYS', 'SER', 'ARG', 'TRP', 'SER', 'GLY', 'SER', 'HIS', 'GLN', …, ...
…,
```

- **models.csv**

```
identifier,structure_file,scores,index
loop_1, ... ,"{'ramachandran_outliers': 0.2857, 'rotamer_outliers': 0.3571, …, 'molprobity_score': 2.72}",1
…,
```

Example: 3IDP





- Proof of concept
- Results not usable as is

Next steps

- **Energy minimisation or short MD to sanitise generated structures**
  - **vacuum or solvated**
  - **possibly already before scoring**
    (`Minimiser` could formally be a `Scorer`)

- **Performance: sequential modelling of isolated segments can be slow**
  - **Naive parallelisation (multiple builders) possible**

- **Consider simultaneous modelling of (coupled) segments**

- **Look into alternative modelling method**
  (AF2 inpainting most promising)