Data curation: Melting point
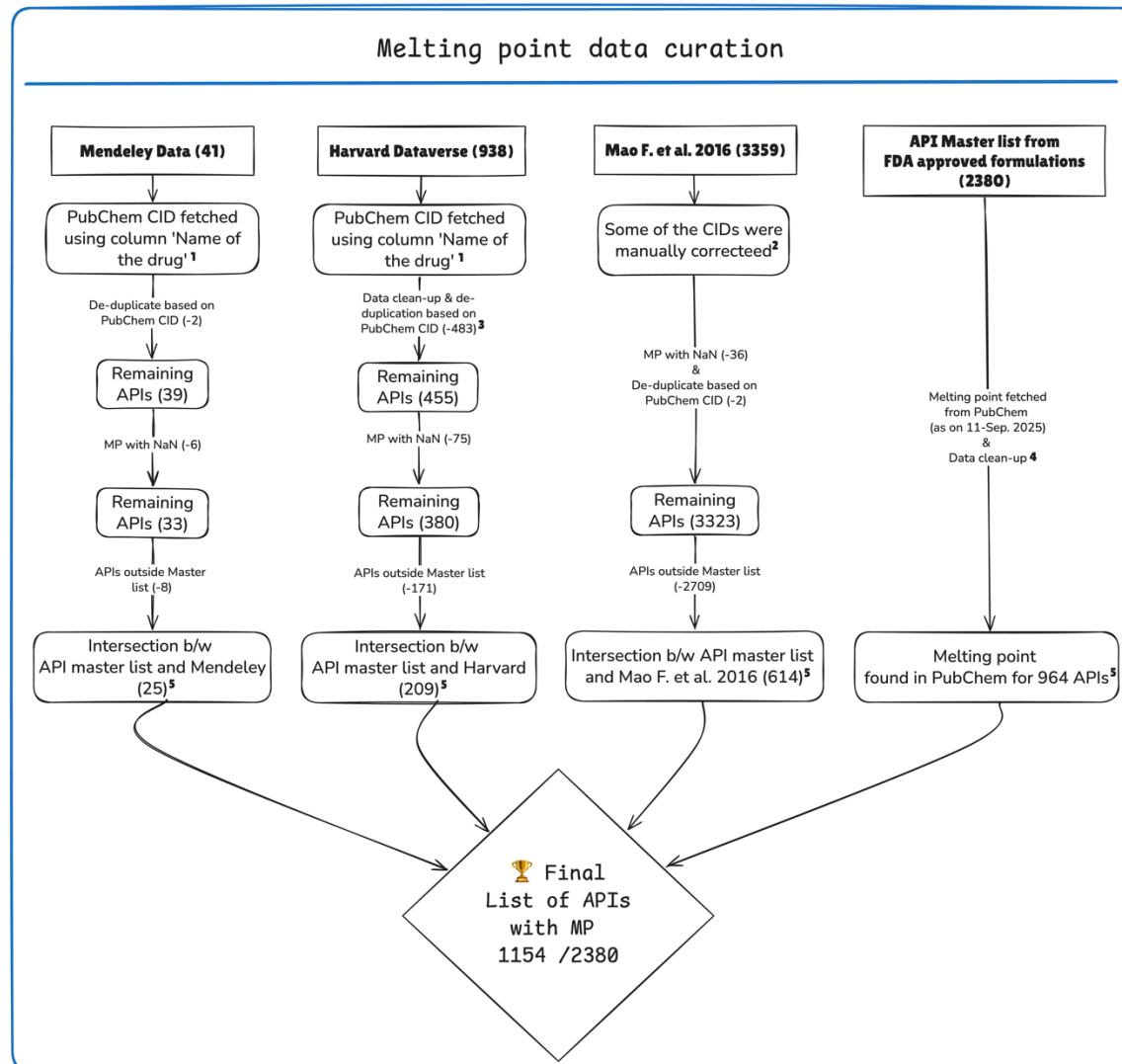
1. **Phase 1:**

   Curation of experimentally determined properties of APIs

   i. Melting point:

   Curated melting point from 4 different public sources *viz.* Mendeley data (25 data points), Harvard dataverse (209 data points), a publication by Mao. F., *et al*. 2016 (614 data points), and PubChem (964 data points).

   The data available in the PubChem is an aggregation of multiple sources e.g. for aspirin MP data is gathered from 6 different sources along with references.

Data curation: Melting point

Melting point data curation

Mendeley Data (41)
↓
PubChem CID fetched using column 'Name of the drug'[1]
↓ De-duplicate based on PubChem CID (-2)
Remaining APIs (39)
↓ MP with NaN (-6)
Remaining APIs (33)
↓ APIs outside Master list (-8)
Intersection b/w API master list and Mendeley (25)[5]

Harvard Dataverse (938)
↓
PubChem CID fetched using column 'Name of the drug'[1]
↓ Data clean-up & de-duplication based on PubChem CID (-483)[3]
Remaining APIs (455)
↓ MP with NaN (-75)
Remaining APIs (380)
↓ APIs outside Master list (-171)
Intersection b/w API master list and Harvard (209)[5]

Mao F. et al. 2016 (3359)
↓
Some of the CIDs were manually correcteed[2]
↓ MP with NaN (-36) & De-duplicate based on PubChem CID (-2)
Remaining APIs (3323)
↓ APIs outside Master list (-2709)
Intersection b/w API master list and Mao F. et al. 2016 (614)[5]

API Master list from FDA approved formulations (2380)
↓
Melting point fetched from PubChem (as on 11-Sep. 2025) & Data clean-up [4]
↓
Melting point found in PubChem for 964 APIs[5]

🏆 Final List of APIs with MP 1154 /2380

Notes:

1 = The original dataset does not contain PubChem CIDs. The CIDs were fetched using API names from the respective columns. Using **PUG REST API**.

2 = Some CIDs were incorrect and hence were manually corrected, and missing CIDs were fetched from PubChem using name of API

3 = Data clean-up was done to remove entries with "#VALUE!", "Not specified", absence of API names etc."

4 = Data cleanup was challenging, used Gemini 2.5 Pro LLM and designed 24 different prompts

5 = PubChem CIDs used to match with API of Master list (2380)