GBstudio

A new software GBstudio was developed for generating atomic coordinates in periodic grain boundary models composed two crystals. It was designed for modeling grain boundary structures in various geometries including coincident-site-lattice (CSL), tilt, and twist boundaries in easy and systematic ways. By this software, CSL boundaries of cubic crystals up to $\Sigma 99$ can be constructed by selecting a few parameters in the candidate lists. Tilt and twist boundaries on representative rotation axes can also be generated in a similar way for cubic and non-cubic crystals. An editing menu is implemented to modify inappropriate atomic configuration at the boundary. The software is distributed via the Internet as a Java applet usable on web browsers.

The data in that CSV file represents the **statistical summary of grain boundary (GB) segregation energy distributions** for 260 different binary alloys, as calculated in the Wagih et al. paper.

Instead of a single value, the authors found that the segregation energy for a given alloy forms a "spectrum" (a probability distribution) across all the different sites in a grain boundary.

The CSV file contains the parameters that mathematically define these distributions, which are plotted in Supplementary Figures 2-20.

Column Breakdown

Here's what each column in your wagih_2020_data.csv file means:

- **Solvent**: This is the main, or "host," metal of the alloy (e.g., Al, Ni, Cu).
- **Solute**: This is the "impurity" or alloying element that is dissolved in the solvent (e.g., Mg, Zr, Au).
- **Potential_ID**: This is the specific interatomic potential model used for the computer simulation. The paper's first figure shows that for many alloy pairs, multiple simulation models exist. This ID (e.g., 1, 10, 39) corresponds to the superscript number on the plots (like Al^{1} or Cu^{10}).
- **Mu (μ)**: This is the **location parameter**, which is similar to the **average segregation energy** in kJ/mol. A negative value (e.g., μ=-3) means segregation is favorable on average.
- **Sigma (σ)**: This is the **scale parameter**, which is similar to the **standard deviation**. It measures the *width* or *spread* of the energy distribution. A large Sigma (e.g., σ=31) means there is a very wide range of different segregation energies possible in the grain boundaries.
- **Alpha (α)**: This is the **skewness parameter**. It measures how asymmetric the distribution is.
    - If α = 0, the distribution is symmetric (a perfect bell curve).
    - If α < 0 (negative), the distribution has a long "tail" towards more negative (stronger) segregation energies.
- **R_Squared (R^2)**: This is the **"goodness-of-fit"**. It shows how well this 3-parameter model (μ, σ, α) fits the authors' raw simulation data. A value of 1.00 is a perfect fit, and 0.98 is a very good fit.

The data in that CSV file represents the **statistical summary of grain boundary (GB) segregation energy distributions** for 260 different binary alloys, as calculated in the Wagih et al. paper.

Instead of a single value, the authors found that the segregation energy for a given alloy forms a "spectrum" (a probability distribution) across all the different sites in a grain boundary.

The CSV file contains the parameters that mathematically define these distributions, which are plotted in Supplementary Figures 2-20.

Column Breakdown

Here's what each column in your wagih_2020_data.csv file means:

- **Solvent**: This is the main, or "host," metal of the alloy (e.g., Al, Ni, Cu).
- **Solute**: This is the "impurity" or alloying element that is dissolved in the solvent (e.g., Mg, Zr, Au).
- **Potential_ID**: This is the specific interatomic potential model used for the computer simulation. The paper's first figure shows that for many alloy pairs, multiple simulation models exist. This ID (e.g., 1, 10, 39) corresponds to the superscript number on the plots (like Al^{1} or Cu^{10}).
- **Mu (μ)**: This is the **location parameter**, which is similar to the **average segregation energy** in kJ/mol. A negative value (e.g., μ=-3) means segregation is favorable on average.
- **Sigma (σ)**: This is the **scale parameter**, which is similar to the **standard deviation**. It measures the *width* or *spread* of the energy distribution. A large Sigma (e.g., σ=31) means there is a very wide range of different segregation energies possible in the grain boundaries.
- **Alpha (α)**: This is the **skewness parameter**. It measures how asymmetric the distribution is.
    - If $\alpha = 0$, the distribution is symmetric (a perfect bell curve).
    - If $\alpha < 0$ (negative), the distribution has a long "tail" towards more negative (stronger) segregation energies.
- **R_Squared (R^2)**: This is the **"goodness-of-fit"**. It shows how well this 3-parameter model (μ, σ, α) fits the authors' raw simulation data. A value of 1.00 is a perfect fit, and 0.98 is a very good fit.

What this means for your project

This file is a "summary" dataset. It's different from your JARVIS dataset, which likely contains *per-site* energies.

You can use this wagih_2020_data.csv file in two ways:

1. **As a target:** Train a machine learning model to predict the Mu, Sigma, and Alpha parameters for any given alloy.
2. **As a data generator:** Use these parameters to sample from the statistical distributions and generate a massive "virtual" dataset of segregation energies to train your neural networks.

Based on the dataset I provided from the Wagih et al. paper, here is the cleaning and preprocessing required.

The short answer is: **very little data *cleaning* is needed, but significant data *preprocessing* is essential.**

The data you've just created is already in excellent shape because it was transcribed directly from the paper's figures.

Data Cleaning (What to check for)

1. **Missing Values:** Unlikely. Every plot in the paper had values for Mu, Sigma, and Alpha. A quick check for null or NaN values is still good practice, but you likely won't find any.
2. **Duplicates:** Not an issue. The combination of Solvent, Solute, and Potential_ID is unique for each entry.
3. **Outliers / Quality Control:** This is the *only* cleaning step you might consider.
    - The **R_Squared** column tells you how well the (Alpha, Mu, Sigma) parameters fit the simulation data.
    - Most values are high (e.g., 0.98 to 1.00).
    - You could, for example, filter out rows where the fit is "poor." The Cr-Fe entry (ID 30) has an R^{2} of 0.80[1], which is the lowest. You might decide to remove any entry with R^{2} < 0.90 to ensure your model only trains on high-quality data.

Data Preprocessing (What you *must* do)

This is the most critical part and aligns perfectly with **Phase 3: Feature Engineering** of your project plan[2]. Your machine learning model cannot understand text like "Al" or "Mg". You must convert them into numbers.

1. Categorical Features (Solvent, Solute)

This is your main task. These two columns are your inputs (features). You need to convert these element symbols into numerical vectors. Your project plan specifically tells you how to do this:

- **Element Property Vectors:** You will replace "Al" (as the solvent) and "Mg" (as the solute) with vectors of their physical properties.
- **Matminer:** Your plan mentions using matminer composition featurizers[3]. This is a Python library that can automatically convert "Al" into a list of features like:
    - Atomic radius
    - Electronegativity
    - Number of valence electrons
    - Heat of formation
- **Feature Engineering:** You will also likely create *new* features, such as the *difference* between the solvent's atomic radius and the solute's atomic radius.

2. Numerical Targets (Mu, Sigma, Alpha)

These three columns are your outputs (targets). Your model will try to predict these values.

- **No Scaling (for prediction):** Since these are your final targets, you typically *do not* scale or normalize them. You want the model to predict the energy in its real units (kJ/mol).
- **Data Type:** You just need to ensure these columns are read by pandas as numeric (float) types.

3. Identifiers (Potential_ID)

This column is just an identifier to distinguish between different models for the same alloy pair (e.g., the 5 different models for Ag-Cu 4444). You should not use this column as a feature to train your model, as it's just a label with no physical meaning.

Summary: Your Next Step

Your next step is to write a Python script that loads this wagih_2020_data.csv and uses a library like **Matminer** to convert the Solvent and Solute columns into a rich set of numerical features.

The μ (mu) value in your CSV file represents the **statistical mean (or average) segregation energy** for a specific alloy combination, measured in **kilojoules per mole (kJ/mol)**[111111111].

In your project, this is one of the most important values you will be working with. Here's what its value tells you:

Physical Interpretation

Segregation energy is the energy difference between a solute atom residing at a grain boundary (GB) versus residing inside the "bulk" (the middle of a crystal grain). The μ value tells you the *average* tendency of that solute.

- If μ is Negative (e.g., -50 kJ/mol):

This means the solute atom is, on average, more stable (at a lower energy) at the grain boundary. This creates a strong driving force for the solute to accumulate at the GBs. This is favorable segregation.

  - **Example:** For the Mo (solvent) - Pb (solute) alloy, μ = -52[2]. This indicates that Lead (Pb) atoms will strongly segregate to the grain boundaries in Molybdenum (Mo).

- If μ is Positive (e.g., +50 kJ/mol):

This means the solute atom is, on average, less stable (at a higher energy) at the grain boundary. The solute atoms will actively avoid the GBs and prefer to stay in the bulk. This is unfavorable segregation (or depletion).

  - **Example:** For the Al (solvent) - Nb (solute) alloy, μ = 72[3]. This indicates that Niobium (Nb) atoms will be repelled from the grain boundaries in Aluminum (Al).

- If μ is Near Zero (e.g., -2 to +2 kJ/mol):

This means there is, on average, no significant energy difference between the bulk and the grain boundary. There is no strong driving force for the solute to either accumulate at or avoid the GBs.

- **Example:** For the Ag (solvent) - Au (solute) alloy, one potential gives μ = 0[4].

Why is it a "Mean" (μ)?

The Wagih et al. paper treats segregation energy not as a single value but as a **distribution (or "spectrum")**[5]. This is because a grain boundary is a complex interface with thousands of different atomic sites, each with a slightly different energy.

The μ value is the **location parameter** (or mean) of this entire distribution[6], while σ (sigma) describes its width and α (alpha) describes its skewness. For your project, μ is the primary target for predicting the *average* segregation behavior.

https://www.jstage.jst.go.jp/article/matertrans/47/11/47_11_2706/_article