Lessons learned from database developments

ATLASx database incorporates millions of compounds and reactions. Dealing with such a large and diverse set of compounds and reactions has important technical details. it is essential to check the data from all possible imaginable angles to ensure quality. Here, we explain some of the important points we realized:

- Each aromatic molecule has different kekuléd representations. Depending on the structure of a molecule the combination of atom-bonds configurations in aromatic rings can result to hundreds of kekuléd representations. **Cheminformatic investigations are sensitive to kekuléd forms** and depending on which forms is considered the results may vary. To predict all possible reactions in ATLASx, we took into account all possible kekuléd forms of the molecules. Analyzing millions of kekuléd forms is not feasible without proper infrastructure. In addition, a specific form of reactions that happens between kekuléd representation of the same molecule requires extra attention. Basically, if two kekuléd forms of one compound are reacting with each other, reaction will be formulated as:

  Compound_A_ kekuléd form_1 + Compound_A_ kekuléd form_2 +… $\rightarrow$ products

  Which should be modified as:

  (2) Compound A +… $\rightarrow$ products

- Salt compounds consist of more than one component and each component reacts differently. So, **each component of a salt should be analyzed separately.** Otherwise, they create unbalanced reactions.

- To make sure reactions are atom balanced and have all the required information, controlling quality of generated data is essential. In ATLASx, each new reaction is tested with a **quality control function before insertion to the database**.

- Finally, to avoid incomplete import of data, which can be resulted from instability of internet connections, we suggest **keeping track of transactions**. In case of internet instability this information helps to restart connection and resend the last transactions.