The Journal of Open Source Software

GCIdentifier.jl: A Julia package for identifying molecular fragments from SMILES

Pierre J. Walker [1,2¶], **Andrés Riedemann**[3¶], **and Zhen-Gang Wang** [1]

1 Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, United States **2** Department of Chemical Engineering, Imperial College, London SW7 2AZ, United Kingdom **3** Departamento de Ingeniería Química, Universidad de Concepción, Concepción 4030000, Chile ¶ Corresponding author

DOI: 10.21105/joss.06453

Software

- Review
- Repository
- Archive

Editor: Bonan Zhu

Reviewers:

- @Arrondissement5etDemi
- @mjohnson541
- @moyner

Submitted: 27 February 2024
Published: 04 April 2024

Summary

GCIdentifier.jl is an open-source toolkit for the automatic identification of group fragments based on the name of a molecule or its SMILES. Obtaining chemical properties of species, such as heat capacities (Benson & Buss, 1958) or solvation free energies (Platts et al., 2000), will typically involve a set of parameters that represent a given species. For example, ideal isobaric heat capacities over a range of temperature of a pure component can be obtained using Reid polynomials with just four parameters (a, b, c and d). Unfortunately, in this case, the parameters obtained are only applicable to a specific species and cannot be transferred to others (i.e. the a, b, c, d parameters for water cannot then be used to model ibuprofen). A solution to this would be to split a set of molecules with similar chemical structures into moieties, known as groups, each of which will have their own parameters associated with them and adjust these parameters against experimental data for all of these molecules. The combination of these groups (and their associated parameters) can then be used to predict the properties of ibuprofen. In the case of the Joback method(Joback & Reid, 1987), the Reid polynomial parameters can be obtained by summing over the group-specific parameters (a_i, b_i, c_i and d_i) weighted by the occurrence of those groups in a species. The benefit of such approaches is that these groups can be combined many different ways such that they represent a larger variety of molecules. This type of approach is known as group contribution, where many examples of such approaches exist (Chung et al., 2022; Papaioannou et al., 2014; Walker & Haslam, 2020; Weidlich & Gmehling, 1987) which can be used to predict a range of properties such as pharmaceutical solubilities(Wehbe et al., 2022), interfacial tensions(Rehner et al., 2021) and thermal conductivites(Hopp & Gross, 2019). An example of this process is shown in figure 1.

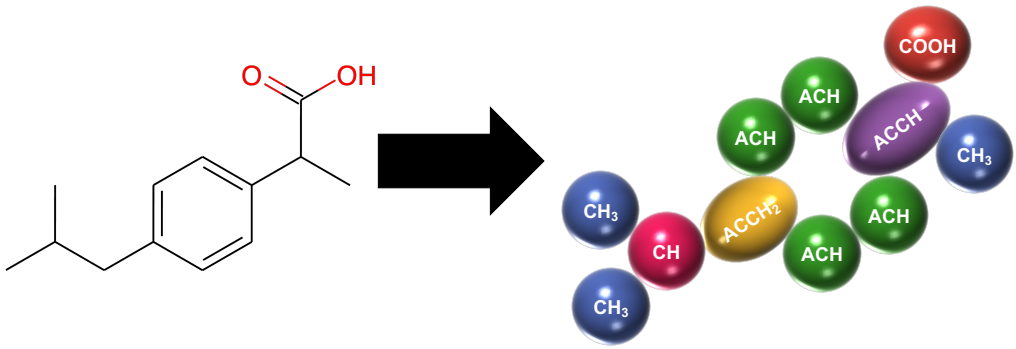

Figure 1: Fragmentation of ibuprofen into UNIFAC groups.

Unfortunately, the challenge with using group-contribution approaches is the assignment of the groups to represent a given species. While this assignment can be done manually, it is more convenient and, as discussed later, efficient to automate this process. Indeed, this is the exact objective of GCIdentifier. By simply feeding a species name or SMILES, along with the group-contribution approach one wishes to use, the group assignment is done automatically:

```
using GCIdentifier, ChemicalIdentifiers

groups = get_groups_from_name("ibuprofen", UNIFACGroups)
```

The output from this function can then be used in other packages, such as Clapeyron (Walker et al., 2022), to obtain chemical properties.

Statement of need

Group-contribution approaches are vital when it comes to computer-aided molecular design (CAMD) of, for example, novel refrigerants (Sahinidis et al., 2003) or in drug discovery(Hou et al., 2004). Here, the assignment of groups must be done thousands of times and, in some cases, for rather complex molecules. This is the primary motivator for the development of GCIdentifier. While other packages(Degen et al., 2008; Liu et al., 2017; Müller, 2019) with similar functionalities have been developed in other languages, GCIdentifier.jl stands apart for multiple reasons.

GCIdentifier.jl is the first of such packages to be compatible with multiple group-contribution approaches, such as UNIFAC and SAFT-γ Mie. By standardising the representation of groups using SMARTS and leveraging the powerful MolecularGraph (Matsuoka et al., 2024) package, our group-identification code can be used with any existing group-contribution thermodynamic model. This extends to group-contribution approaches which require information about the connectivity between groups (Sauer et al., 2014) where, by simply specifying `connectivity=true` within the `get_groups_from_name` function, the connectivity matrix between groups will automatically be generated.

While packages in other languages are able to generate groups from *existing* group databases, GCIdentifier.jl is able to systematically propose *new* groups for a given molecule. Consider a case where an existing group-contribution framework is unable to cover all atoms present in a molecule. GCIdentifier.jl is able to consider these un-represented atoms and propose a list of new groups. From this list, users will be able to determine which groups they should obtain new parameters for. In the extreme case where we wish to generate a list of all possible groups that represent a molecule, GCIdentifier.jl will automatically split the molecule into groups, from

which either the user or a set of built-in heuristics can then decide which set best represent the molecule.

These two features present within GCIdentifier.jl have potential applications beyond thermodynamic modelling, such as the development of molecular dynamics forcefields which could be integrated into packages such as Molly (Greener, 2023).

Acknowledgments

Z-G.W. acknowledges funding from Hong Kong Quantum AI Lab, AIR@InnoHK of the Hong Kong Government.

References

Benson, S. W., & Buss, J. H. (1958). Additivity rules for the estimation of molecular properties. Thermodynamic properties. *The Journal of Chemical Physics*, *29*(3), 546–572. https://doi.org/10.1063/1.1744539

Chung, Y., Vermeire, F. H., Wu, H., Walker, P., Abraham, M. H., & Green, W. H. (2022). Group contribution and machine learning approaches to predict abraham solute parameters, solvation free energy, and solvation enthalpy. *J. Chem. Inf. Model.*, *62*(3), 433–446. https://doi.org/10.1021/acs.jcim.1c01103

Degen, J., Wegscheid-Gerlach, C., Zaliani, A., & Rarey, M. (2008). On the art of compiling and using 'drug-like' chemical fragment spaces. *ChemMedChem*, *3*(10), 1503–1507. https://doi.org/10.1002/cmdc.200800178

Greener, J. G. (2023). *Differentiable simulation to develop molecular dynamics force fields for disordered proteins*. bioRxiv. https://doi.org/10.1101/2023.08.29.555352

Hopp, M., & Gross, J. (2019). Thermal conductivity from entropy scaling: A group-contribution method. *Industrial & Engineering Chemistry Research*, *58*(44), 20441–20449. https://doi.org/10.1021/acs.iecr.9b04289

Hou, T. J., Xia, K., Zhang, W., & Xu, X. J. (2004). ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *Journal of Chemical Information and Computer Sciences*, *44*(1), 266–275. https://doi.org/10.1021/ci034184n

Joback, K. G., & Reid, R. C. (1987). Estimation of pure-component properties from group-contributions. *Chem. Eng. Commun.*, *57*(1), 233–243. https://doi.org/10.1080/00986448708960487

Liu, T., Naderi, M., Alvin, C., Mukhopadhyay, S., & Brylinski, M. (2017). Break down in order to build up: Decomposing small molecules for fragment-based drug design with eMolFrag. *Journal of Chemical Information and Modeling*, *57*(4), 627–631. https://doi.org/10.1021/acs.jcim.6b00596

Matsuoka, S., Holy, T., hhaensel, Henle, A., TagBot, J., Richard, McGrath, T., & Box, W. (2024). *Mojaie/MolecularGraph.jl: v0.16.0* (Version v0.16.0). Zenodo. https://doi.org/10.5281/zenodo.10478701

Müller, S. (2019). Flexible heuristic algorithm for automatic molecule fragmentation: Application to the UNIFAC group contribution model. *Journal of Cheminformatics*, *11*(1), 57. https://doi.org/10.1186/s13321-019-0382-3

Papaioannou, V., Lafitte, T., Avendaño, C., Adjiman, C. S., Jackson, G., Müller, E. A., & Galindo, A. (2014). Group contribution methodology based on the statistical associating fluid theory for heteronuclear molecules formed from mie segments. *J. Chem. Phys.*, *140*(5), 054107. https://doi.org/10.1063/1.4851455

Platts, J. A., Abraham, M. H., Butina, D., & Hersey, A. (2000). Estimation of molecular linear free energy relationship descriptors by a group contribution approach. 2. Prediction of partition coefficients. *J. Chem. Inf. Model.*, *40*(1), 71–80. https://doi.org/10.1021/ci990427t

Rehner, P., Bursik, B., & Gross, J. (2021). Surfactant modeling using classical density functional theory and a group contribution PC-SAFT approach. *Industrial & Engineering Chemistry Research*, *60*(19), 7111–7123. https://doi.org/10.1021/acs.iecr.1c00169

Sahinidis, N. V., Tawarmalani, M., & Yu, M. (2003). Design of alternative refrigerants via global optimization. *AIChE Journal*, *49*(7), 1761–1775. https://doi.org/10.1002/aic.690490714

Sauer, E., Stavrou, M., & Gross, J. (2014). Comparison between a homo- and a heterosegmented group contribution approach based on the perturbed-chain polar statistical associating fluid theory equation of state. *Ind. Eng. Chem. Res.*, *53*(38), 14854–14864. https://doi.org/10.1021/ie502203w

Walker, P. J., & Haslam, A. J. (2020). A new predictive group-contribution ideal-heat-capacity model and its influence on second-derivative properties calculated using a free-energy equation of state. *J. Chem. Eng. Data*, *65*(12), 5809–5829. https://doi.org/10.1021/acs.jced.0c00723

Walker, P. J., Yew, H.-W., & Riedemann, A. (2022). Clapeyron.jl: An extensible, open-source fluid thermodynamics toolkit. *Ind. Eng. Chem. Res.*, *61*(20), 7130–7153. https://doi.org/10.1021/acs.iecr.2c00326

Wehbe, M., Haslam, A. J., Jackson, G., & Galindo, A. (2022). Phase behaviour and pH-solubility profile prediction of aqueous buffered solutions of ibuprofen and ketoprofen. *Fluid Phase Equilibria*, *560*, 113504. https://doi.org/10.1016/j.fluid.2022.113504

Weidlich, U., & Gmehling, J. (1987). A modified UNIFAC model. 1. Prediction of VLE, h^E, and γ. *Ind. Eng. Chem. Res.*, *26*(7), 1372–1381. https://doi.org/10.1021/ie00067a018