

SELFIES and the future of molecular string representations

discussion points

<https://accelerationconsortium.substack.com/p/selfies-workshop-aug-13>

Discord for follow-up discussions: <https://discord.gg/M8GsAvbs>
[Paper](#) and [GitHub](#)



UNIVERSITY OF
TORONTO



1) SELFIES for Macromolecules

A recent paper (<https://pubs.acs.org/doi/10.1021/acscentsci.9b00476>) has shown how to represent polymers using SMILES, by extending SMILES to BigSMILES. How can we translate this representation to BigSELFIES?

Kick-off questions:

- What are the difficulties going from molecules to macromolecules, e.g., polymers?
- How is this done so far?
- How does BigSMILES work?
- What are exciting inverse-design tasks for polymers?

2) SELFIES for crystal structures

Kick-off questions:

- How are crystals represented in SMILES or with other (string-based) representations?
- What are the issues with this representation?
- What are interesting inverse-design problems for solid-state material design?
- What are interesting inverse-design problems for crystal design?

3) Nontrivial bonds

SMILES and SELFIES cannot represent certain bonds in a natural way. Examples are Metallocenes or Diborane. How can they be described using strings?

Kick-off questions:

- What is the issue of Metallocenes (Ferrocene) in SMILES?
- What other bonds cannot (naturally) be described in SMILES?
- How could we represent Metallocenes in a more natural way?
- What are interesting inverse-design problems for molecules with nontrivial bonds?

4) reaction SELFIES and SMARTS for SELFIES?

SMILES have been extended to allow for [descriptions of reactions](#)
SELFIES/CGR in similar way as [SMILES/CGR](#).

Kick-off questions:

- a) What possibilities would a completely robust reaction representation open?
- b) What could we do with a transformation SELFIES (reaction template)?
- c) What is a great application of a 100% robust reaction language?
- d) How would it compare to state of the art methods?

5) Human-interpretable Representation

Which representations are easier to learn, to read and to write for humans, SMILES or DeepSMILES or SELFIES?

Kick-off questions:

- a) How could we make a scientific experiment for this?
- b) Is there a difference between experts and non-experts?
- c) Are there scientific protocols to follow for experiments of this sort?

6) Machine-interpretability?

Which representations can be learned and applied easier by AI systems? An examples for unexpected advantage for SELFIES: [Image2String](#) and [String2String](#) translation.

Kick-off questions:

- a) What does it actually mean that a machine understands one representation better than the other?
- b) How can one compare representations?
- c) How could one design new experiments to compare between SMILES and SELFIES in Deep Learning Settings?
- d) How to consider the information content per character and effort to learn to use the representations?

7) String-representations as a programming language: More powerful than “graphs”

One can think about SMILES and SELFIES as a programming language, encoder/decoder are the interpreters (such as interpreter in Python/JavaScript), then a generative model is programming in that language.

Kick-off questions:

- a) How can we show this explicitly, what else can we show in this way, connection to language models?
- b) By definition a general programming language is much more powerful than adjacency matrices that are used in Graph-Based representations. How to explicitly show that?
- c) What chemical structural properties cannot be represented by adjacency matrices?
- d) How does OpenAI's Codex (the AI that writes code)?

8) [coding project] metaSELFIES as an interdisciplinary AI representation

The grammar of SELFIES can be extracted directly from data. This would be possible not only for inverse-design in chemistry, but also in [quantum optics](#) and likely many other fields.

Example: we have some graph-based datasets for chemistry/maybe RNA origamis/some quantum system. So we have one VAE, one "meta-SELFIES" (that from the dataset creates the chemistry-SELFIES, the RNA-origami-SELFIES,...), and can do highly cross-domain learning.

The basis of the algorithm is described in the [SI of the SELFIES paper](#).

Kick-off questions:

- a) How to use an interdisciplinary robust representation for AI?
- b) Which disciplines (except of chemistry and quantum optics) could be target?
- c) What are exciting questions for an interdisciplinary AI?