

## Background

We have developed ChemHits (<http://sabio.h-its.org/chemHits/>), an application which detects and matches synonymic names of chemical compounds and thereby facilitates the bundling of corresponding data referring to the same compound, but described with different names. The tool that we have developed is based on natural language processing (NLP) methods and applies rules to systematically normalize chemical compound names. Subsequently, matching of synonymous names is achieved by comparison of the normalized name forms. The tool is capable of normalizing a given name of a chemical compound and matching it against names in (bio-)chemical databases, like SABIO-RK, PubChem, ChEBI or KEGG, even when there is no exact name-to-name-match.

## Terminology of Chemical Compounds

### Synonymous notations of chemical compounds

#### Trivial name and systematic chemical description

Valproic acid = 2-Propylpentanoic acid

#### Different parts of the molecule could be considered as lead structure

Acetylphenol = Phenylacetate

#### Aberrant order of the substituents of a lead structure (prefixes)

2-Amino-6-methyl-4-pyrimidol = 6-Methyl-2-amino-4-pyrimidol

#### Description of substituents as prefix (like amino-) or suffix (like -amine)

2-Aminopropane = Propan-2-amine

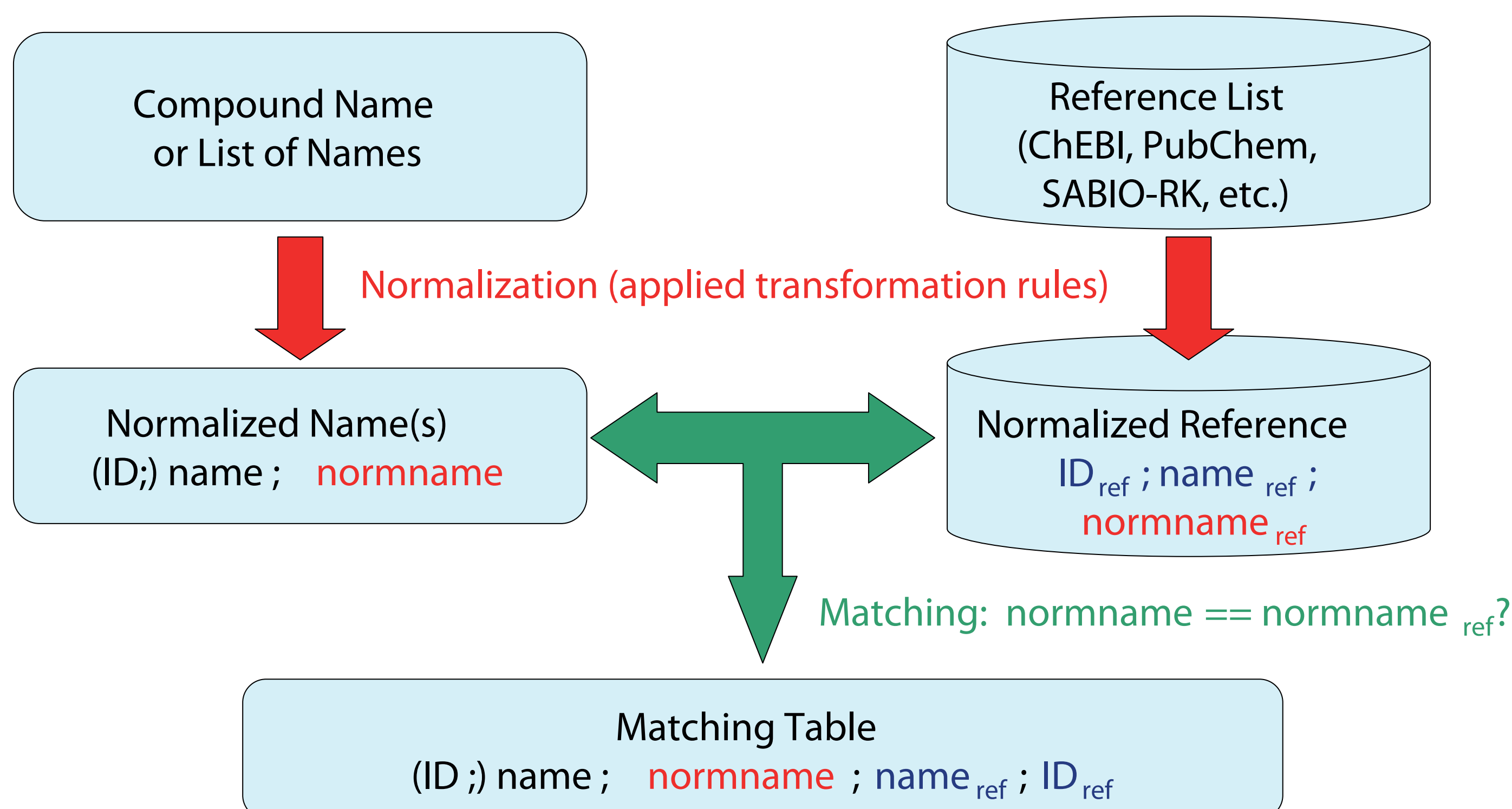
2-Methylpropan-2-ol = 2-Hydroxy-2-methyl-propane

#### Different nomenclature systems (e.g. aberrant order of the morphemes)

2-Amino-6-methyl-4-pyrimidol = 2-Amino-6-methylpyrimidin-4-ol

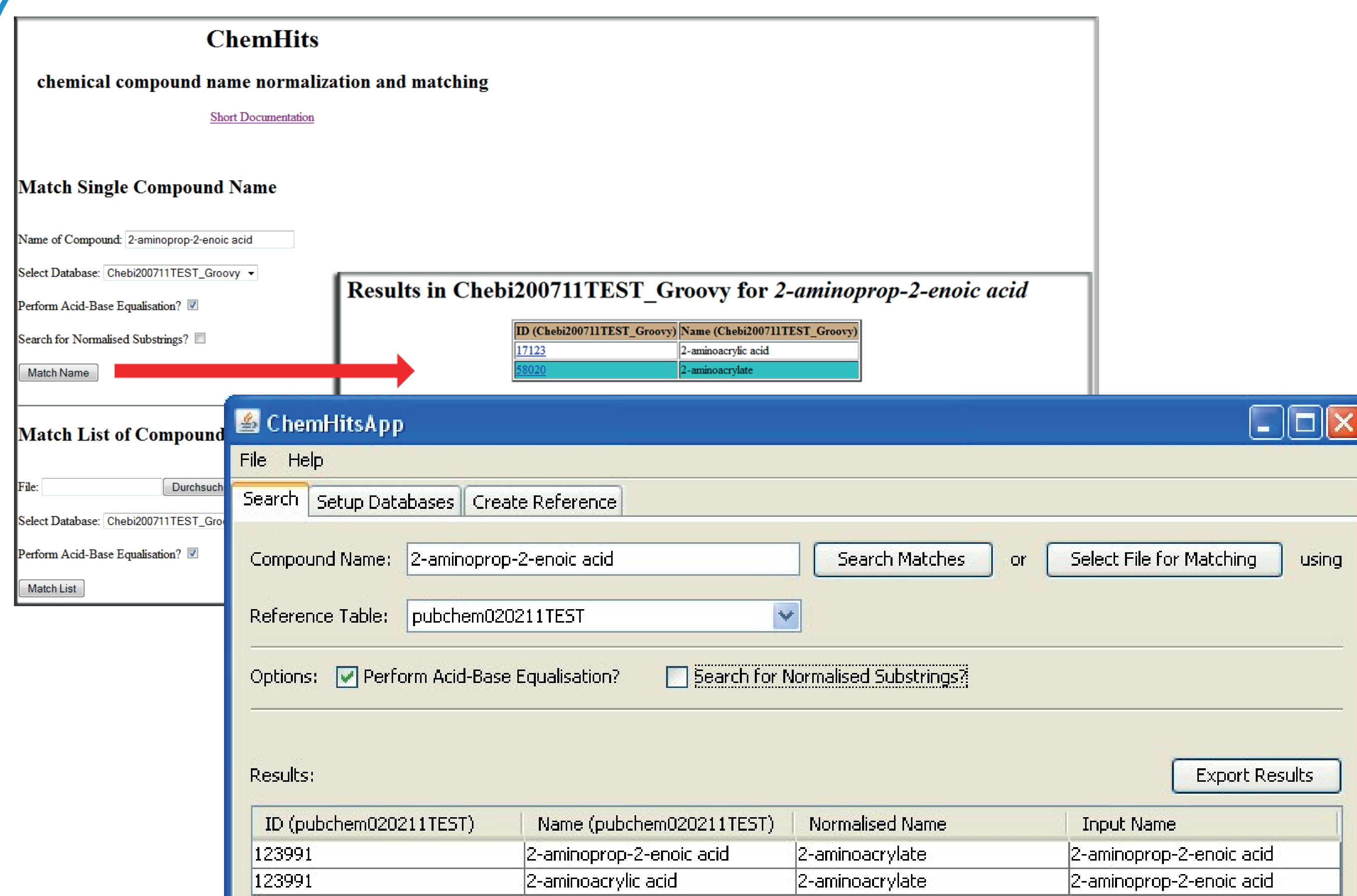
A chemical compound can have many different names - trivial, as well as systematic names. Hence, the identification of a chemical compound solely based on its name requires comprehensive chemical knowledge and often extensive searches in chemical databases. However, this identification is crucial for the integration of biochemical data e.g. for the setup of biochemical models based on published data. As many publications exclusively describe a chemical compound by its name the matching of these diverging notations can be tedious.

## Normalization of Chemical Compound Names



The tool that we have developed applies transformation rules to systematically normalize the notation of chemical compound names [1]. Subsequently, matching of synonymous names is achieved by comparison of the normalized name forms. The normalization rules include, among others, reordering of substituent descriptions in the name and replacement of synonymous name constituents (e.g. equivalent trivial names). Matching of conjugated acid-base pairs is optional for biochemicals.

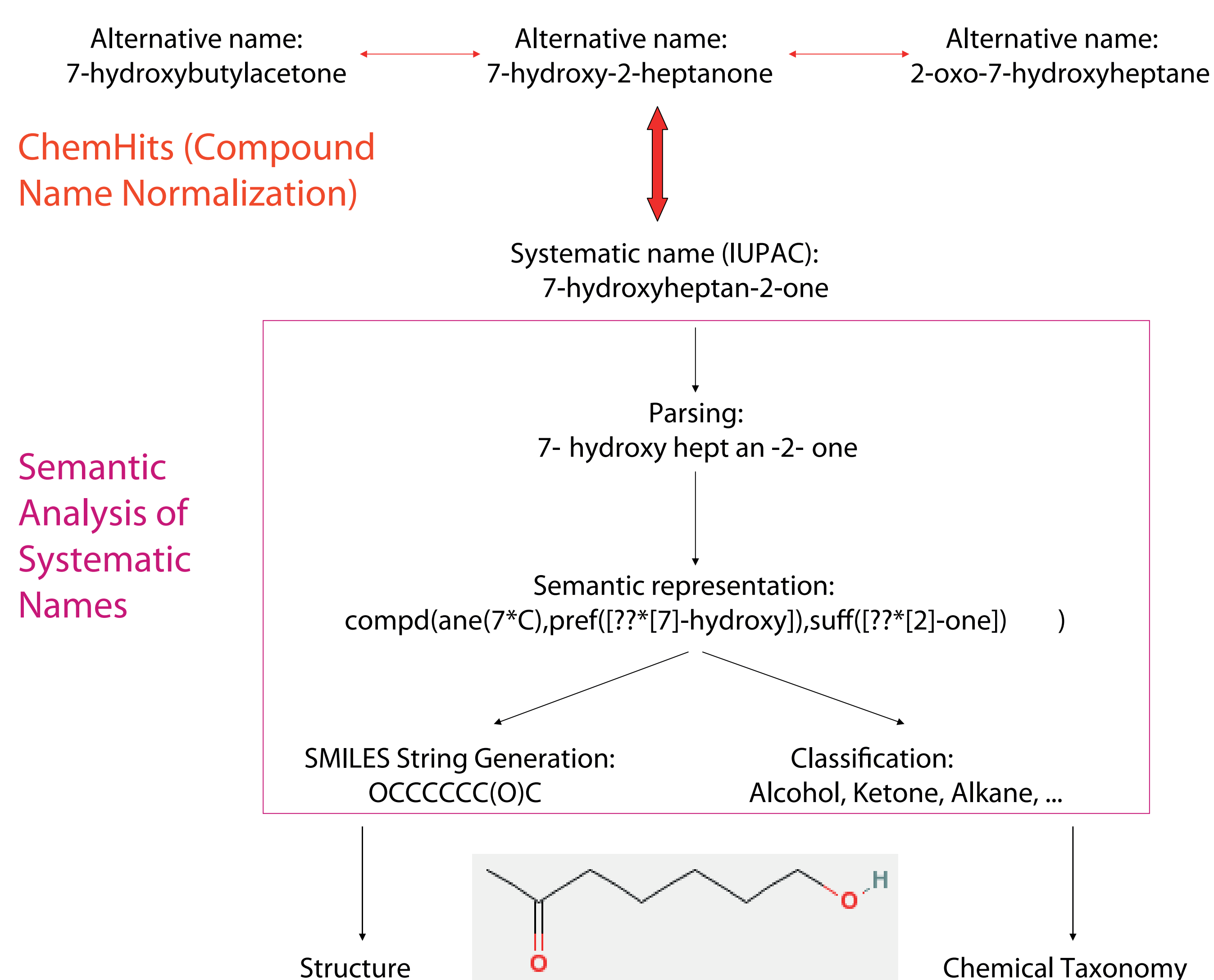
## Chemical Compound Name Matching



ID (pubchem020211TEEST)	Name (pubchem020211TEEST)	Normalised Name	Input Name
123991	2-aminoprop-2-enoic acid	2-aminoacrylate	2-aminoprop-2-enoic acid
123991	2-aminoacrylic acid	2-aminoacrylate	2-aminoprop-2-enoic acid

The tool is capable of normalizing a given name of a chemical compound and matching it against names in (bio-)chemical databases, like ChEBI (<http://www.ebi.ac.uk/chebi/>), KEGG COMPOUND (<http://www.genome.jp/kegg/compound/>), SABIO-RK (<http://sabio.h-its.org/>), or PubChem (<http://pubchem.ncbi.nlm.nih.gov/>), even when there is no exact name-to-name-match. The tool is also able to match a complete list of compound names against these databases which makes it useful for the automatic cross-annotation of chemical data in databases. We offer a web-based service with pre-normalized reference lists (upper screenshot) and a tool that can be installed and run locally with any given reference list (lower screenshot).

## The Future: From Name to Structure Combining Name Matching and Semantic Analysis



After normalization, synonymous notations could potentially be matched to the corresponding systematic name as defined by the International Union of Pure and Applied Chemistry (IUPAC). When combined with our approach to construct chemical structures from systematic names [2, 3], notations could be translated into a chemical structure (SMILES) and classified by functional groups [4], resulting in the unambiguous identification of these compounds.

## References

- Engelken H, Golebiewski M, Bittkowski M, Hamm F, Saric J, Wittig U, Müller W, Reyle U, Rojas I: „Flache und semantische Verarbeitung von Namen biochemischer Verbindungen“, in Stefan Fischer, Erik Maehle, Rüdiger Reischuk: INFORMATIK 2009 - Im Focus das Leben, Beiträge der 39. Jahrestagung der Gesellschaft für Informatik e.V. (GI), Lübeck, Germany, 28. September - 02. Oktober 2009, GI-Edition - Lecture Notes in Informatics (LNI), P-154
- Henriette Engelken: „A System for Semantic Analysis of Chemical Compound Names“, in Proceedings of the ACL-IJCNLP 2009 Student Research Workshop, Singapore, August 2-7, 2009, pages 36-44. Association for Computational Linguistics.
- Kremer G, Anstein S, Reyle U: „Analysing and Classifying Names of Chemical Compounds with CHEMorph“, in S. Ananiadou and J. Fluck: Proceedings of the Second International Symposium on Semantic Mining in Biomedicine: 37-43 (2006)
- Wittig U, Weidemann A, Kania R, Peiss C, Rojas I: „Classification of chemical compounds to support complex queries in a pathway database“, Comparative and Functional Genomics, 5: 156-162 (2004)