

Unique Representations of Elementary Reactions Using InChI-ER

Donald R. Burgess Jr. and Jeffrey A. Manion

Chemical Informatics Research Group
Chemical Sciences Division
National Institute of Standards and Technology
Gaithersburg, Maryland

*Multiagency Coordination Committee for Combustion Research (MACCR)
6th Annual Fuels Research Review*

Argonne National Laboratory
Arlington, Virginia
September 23-26, 2013

InChI-ER

Apply InChI Methodology to Elementary Reactions

- Propose method of extending IUPAC International Chemical Identifier (InChI) to describe and identify elementary reactions.
- Based on existing InChI formalism and concepts, and fully extensible.
- Five layers proposed
 - **Reaction Formula** (A + B → C + D)
 - **Reaction Connectivity** (bonds broken and formed)
 - **Transition State Connectivity** (transition state structure)
 - **Reaction Class** (e.g. bond fission, abstraction)
 - **Chemical Sites** (chemical groups)
- Uniquely identifies elementary reactions (including chemical information).
- Provides chemical information to organize and discover reactions.
- Most layers determined from molecular structures for reactants/products.
- Applied to a representative large hydrocarbon combustion reaction set (based on USC Mech II and JetSurf 2.0).

InChI – Unique Molecular Identifier / Canonical String

- **InChI**: IUPAC International Chemical Identifier (developed at NIST).
- **Widely adopted** as a computer-generated/readable molecular identifier.
- **Line Notation**, String (1-D), Easily Parsed (vs. MOL format: an array)
- Given a specific **molecular structure** – atoms and their connectivities, InChI algorithms returns a **unique identifier**.
- Conversely, an **InChI identifier** can only generate one **unique structure**.
- Molecular structure represented as a **canonical string** – **unique ordering**.
- InChI canonicalization algorithms **serialize molecular structures**.
- Atom labels are **independent of how structure initially drawn**.

Three Steps

- **Normalization** – artifacts removed associated w/ how structure was drawn.
- **Canonicalization** – algorithm employed to number the atoms.
- **Serialization** – atom labels (atom symbols and numbering) are used to generate a string of characters that represent each layer.

InChI is Layered and Extensible

Constitutional Layers

- **Molecular Formula** – Standard Hill-sorted order: C, H, then alphabetically
- **Connectivity** – How heavy (non-hydrogen) atoms are connected
 - Atoms numbered sequentially based on molecular formula layer
 - Ordered using canonicalization algorithm.
- **Hydrogen** – Specifies number of hydrogens attached to each heavy atom

Other Layers

Double Bond Stereochemical, Tetrahedral Enantiomeric, Spatial Inversion Stereoisomers, Charge Distribution, Tautomers, etc, etc

Hierarchical and Extensible

- Layers are hierarchical
- Additional layers provide more information
- New layers (**e.g. InChI-ER**) can be added refining info in previous layers

InChI Constitutional and Double Bond Stereochemical Layers

Molecule	Structure	Formula	Connect	Hydrogen	Stereo
Ethane	CH ₃ CH ₃	C ₂ H ₆	/c1-2	/h1-2 H3	

Ethanol	CH ₃ CH ₂ OH	C ₂ H ₆ O	/c1- 2-3	/h3H,2H2,1H3	
Dimethylether	CH ₃ OCH ₃	C ₂ H ₆ O	/c1- 3-2	/h1-2H3	

(<i>E</i>)-2-Butene	CH ₃ CH=CHCH ₃	C ₄ H ₈	/c1-3-4-2	/h3-4H,1-2H3	/b4- 3+
(<i>Z</i>)-2-Butene	CH ₃ CH=CHCH ₃	C ₄ H ₈	/c1-3-4-2	/h3-4H,1-2H3	/b4- 3-

IUPAC International Chemical Identifier (InChI) a “Structural” identifier, not a “Chemical” identifier

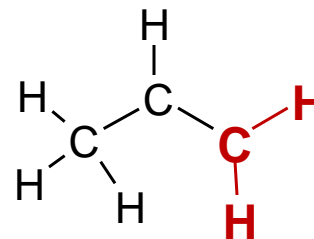
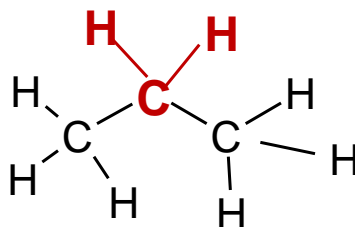
Propane: C₃H₈ /c1-3-2 /h3H₂,1-2H₃

Propene: C₃H₆ /c1-3-2 /h3H,1H₂,2H₃

Different “Chemical” Sites

Propane CH₃**CH₂**CH₃

Propene CH₃CH**CH₂**



Different “Chemical” Reactivities Due to Different Hybridization

Propane CH₃**CH₂**CH₃

–CH₂– group

C-H bond is more reactive

Propene CH₃CH**CH₂**

=CH₂ group

electron rich π bond is more

reactive

Mapping Between InChI and InChI-ER

	InChI Layer		InChI-ER Layer(s)	
1	na	Molecular Formula	/v	Reaction Formula
2	/c	Atom Connectivity	/r	Reaction Connectivity
			/z	Transition State Connectivity
			/k	Reaction Class (reaction functional connectivity)
3	/h	Hydrogen	/w	Chemical Sites

InChI identifies a molecule through the layers

- 1. Molecular Formula** – enumerating (and ordering) the atoms
- 2. Atom Connectivity** – listing the connections between heavy atoms
- 3. Hydrogen** – specifying the number of hydrogen on each site

Examples

Ethanol:	CH ₃ CH ₂ OH	C ₂ H ₆ O	/c1-2-3	/h3H,2H2,1H3
Dimethyl ether:	CH ₃ OCH ₃	C ₂ H ₆ O	/c1-3-2	/h1-2H3

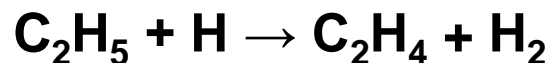
Summary of InChI-ER Layers

	InChI Layer		InChI-ER Layer(s)	
1	na	Molecular Formula	/v	Reaction Formula
2	/c	Atom Connectivity	/r	Reaction Connectivity
			/z	Transition State Connectivity
			/k	Reaction Class (reaction functional connectivity)
3	/h	Hydrogen	/w	Chemical Sites

InChI-ER identifies a reaction through the layers

1. **Reaction Formula** – enumerating/ordering the reactants and products
2. A set of reaction connectivity layers
 - a. **Reaction Connectivity** – “signing” the connections – broken vs. formed
 - b. **Transition State Connectivity** – listing connections in Transition State
 - c. **Reaction Class** (rxn functional connectivity) – characterizing rxn potential
3. **Chemical Sites** – specifying hydrogens AND hybridization, neighbor sites

Identification of Reactants and Products in Not Enough

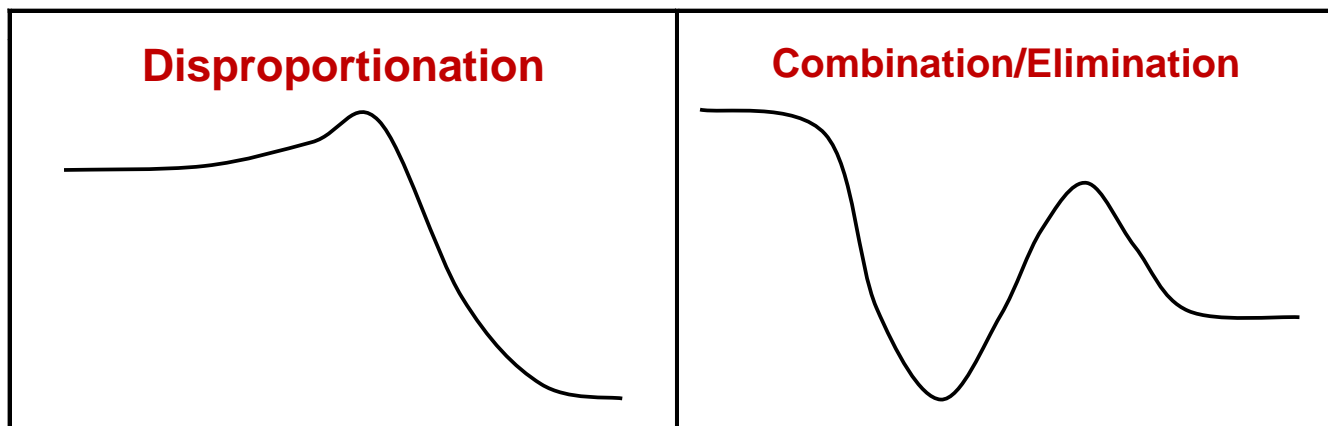


Reaction Type	Transition State	Reaction Layer
Disproportionation	$[\text{CH}_2\text{CH}_2\text{--H--H}]$	/r2-5+1'
Combination/Elimination	$[\text{CH}_3\text{CH}_3]^*$	/r1+1'; 1,2-1',5

Disproportionation: Bimolecular reaction with C-H “abstracted” by H atom

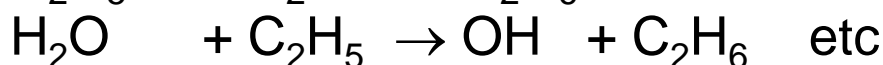
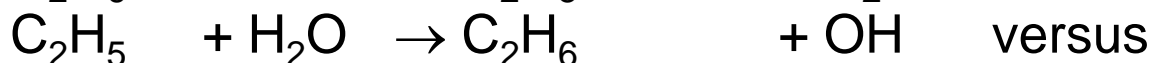
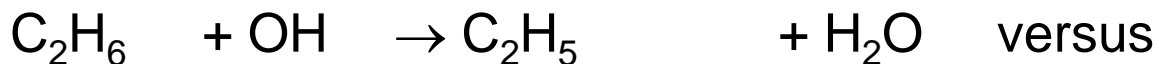
Combination/Elimination: Two step reaction

1. Combination: $\text{C}_2\text{H}_5 + \text{H} \rightarrow \text{CH}_3\text{CH}_3$ (hot)
2. 1,2-Elimination: $\text{CH}_3\text{CH}_3 \rightarrow \text{CH}_2=\text{CH}_2 + \text{H}_2$



Normalization and Ordering of Reactants and Products

How should they be ordered? (we need a standard way)



InChI Solution

(establish rules and stick to them)

InChI-ER Variation

(make them chemical reactivity based)

	<u>Substrate</u>	<u>Reactive</u>	
1.Molecularity	C_2H_6		$\rightarrow \text{CH}_3 + \text{CH}_3$
2.Valence Deficiency	C_2H_6	$+ \text{CH}_3$	$\rightarrow \text{Products}$
3.Bond Deficiency	C_2H_5	$+ \text{C}_2\text{H}_3$	$\rightarrow \text{Products}$
4.Hydrogen Saturation	$\text{CH}_3\text{CH}^*\text{CH}_3$	$+ \text{CH}_3\text{CH}_2\text{CH}_2^*$	$\rightarrow \text{Products}$
5.Chain Length	C_2H_5	$+ \text{CH}_3$	$\rightarrow \text{Products}$
6.Heteroatom Group	CH_3NH_2	$+ \text{CH}_3\text{OH}$	$\rightarrow \text{Products}$
7.Heteroatom Period	CH_3PH_2	$+ \text{CH}_3\text{NH}_2$	$\rightarrow \text{Products}$

Reactants vs Products? (above rules work)

Summary of InChI-ER Reaction Connectivity Layers (3)

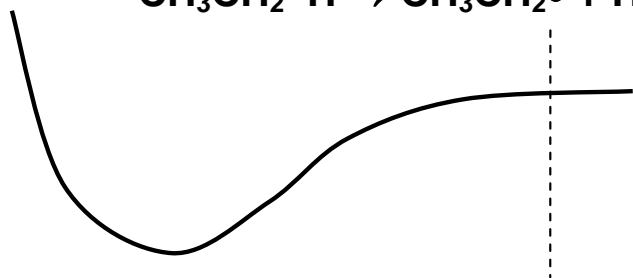
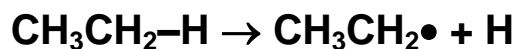
	InChI Layer		InChI-ER Layer(s)	
1	na	Molecular Formula	/v	Reaction Formula
2	/c	Atom Connectivity	/r	Reaction Connectivity
			/z	Transition State Connectivity
			/k	Reaction Class (reaction functional connectivity)
3	/h	Hydrogen	/w	Chemical Sites

InChI-ER identifies a reaction through the layers

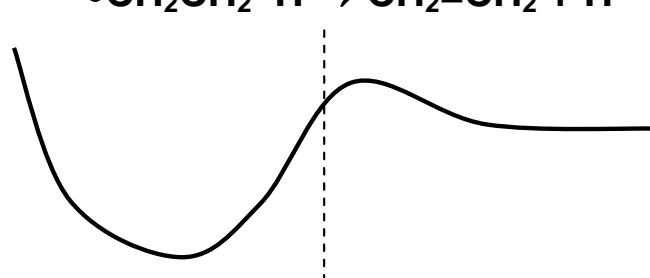
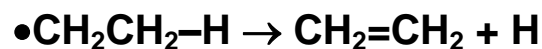
1. **Reaction Formula** – enumerating/ordering the reactants and products
2. A set of reaction connectivity layers
 - a. **Reaction Connectivity** – “signing” the connections – broken vs. formed
 - b. **Transition State Connectivity** – listing connections in Transition State
 - c. **Reaction Class** (rxn functional connectivity) – characterizing rxn potential
3. **Chemical Sites** – specifying hydrogens AND hybridization, neighbor sites

Different Reaction Types/Potential Energy Surfaces

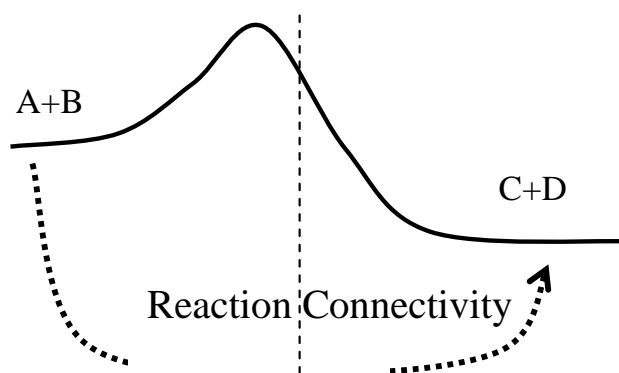
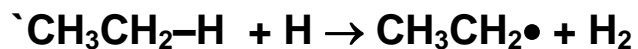
Bond Fission (ethane)



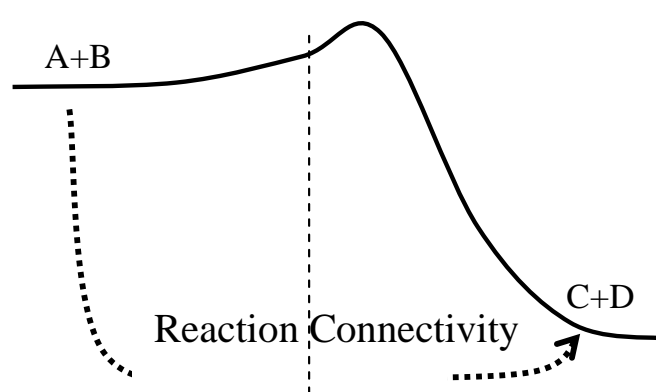
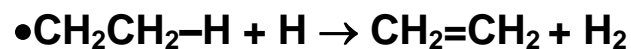
Beta Scission (ethyl)



Abstraction (ethane)



Radical Disproportionation (ethyl)



Simple Example of InChI-ER Layers

CH ₄ + H → CH ₃ + H ₂		
Layer	Description	Example
/v	<u>Reaction Formula</u> Identifies reactants and products	vCH ₄ /h ₁ H ₄ //H ///CH ₃ /h ₁ H ₃ //H ₂ /h ₁ H
/r	<u>Reaction Connectivity</u> Identifies bonds broken and formed	r1-2+1'
/z	<u>Transition State Connectivity</u> Identifies transition state	z1-2-1'
/k	<u>Reaction Class</u> (reaction functional connectivity) Characterizes the type of reaction	R-X+Y//C-H+H (abstraction)
/w	<u>Chemical Sites</u> Identifies sites involved in reaction	CH ₄

More Examples of InChI-ER Layers

Reaction	Reaction Connectivity	Structure
$\text{CH}_4 + \text{H} \rightarrow \text{CH}_3 + \text{H}_2$	/r1-2+1'	

Reaction	InChI Layers	Rxn Layer	Structure
$\text{C}_2\text{H}_6 + \text{OH} \rightarrow \text{C}_2\text{H}_5 + \text{H}_2\text{O}$	C2H6.HO/c1-2; /h1-2H3;1H	/r1-3+1'	
$\text{C}_2\text{H}_6 \rightarrow \text{C}_2\text{H}_4 + \text{H}_2$	C2H6/c1-2 /h1-2H3	/r1,2-3,4	

	Reaction	Transition State
	/r1,2-3,4	/z1-2-4-3-1

Examples of InChI-ER Transition State Layer

Reactants	Products	Trans State	TS Layer
$\text{RCH}_3 + \text{*OH}$	$\rightarrow \text{RCH}_2\text{*} + \text{H}_2\text{O}$	$\text{CH}_3\text{CH}_2\text{--H--OH}$	/z2-4-3
$\text{*CH}_2\text{CH}_3 + \text{OH}$	$\rightarrow \text{CH}_2=\text{CH}_2 + \text{H}_2\text{O}$	$\text{CH}_2\text{CH}_2\text{--H--OH}$	/z2-4-3
$\text{RCH}_2\text{CH}_2\text{R}$	$\rightarrow \text{RCH}=\text{CHR} + \text{H}_2$	$\begin{array}{c} \text{H--H} \\ \text{RCH--CHR} \end{array}$	/z3-4-6-5-3
RCH_3	$\rightarrow \text{RCH:} + \text{H}_2$	$\begin{array}{c} \text{H--H} \\ \text{R--C--H} \end{array}$	/z2-3-4-2
$\text{RCH}_2\text{C*H}_2$	$\rightarrow \text{RCH*CH}_3$	$\begin{array}{c} \text{H} \\ \text{RCH--CH}_2 \end{array}$	/z3-2-4-3

Overview of Reaction Class & Chemical Sites Layers

	InChI Layer		InChI-ER Layer(s)	
1	na	Molecular Formula	/v	Reaction Formula
2	/c	Atom Connectivity	/r	Reaction Connectivity
			/z	Transition State Connectivity
			/k	Reaction Class (reaction functional connectivity)
3	/h	Hydrogen	/w	Chemical Sites

InChI-ER identifies a reaction through the layers

1. Reaction Formula – enumerating/ordering the reactants and products

2. A set of reaction connectivity layers

a. Reaction Connectivity – “signing” the connections – broken vs. formed

b. Transition State Connectivity – listing connections in Transition State

c. Reaction Class (rxn functional connectivity) – characterizing rxn potential

3. Chemical Sites – specifying hydrogens AND hybridization, neighbor sites

Examples of Reaction Class & Chemical Sites Layers

Reaction	Reaction Class	Chemical Sites
propane + OH \rightarrow 1-propyl + H ₂ O	/kC-H+O	/w CH3 ;OH
propane + OH \rightarrow 2-propyl + H ₂ O	/kC-H+O	/w CH2 ;OH
C ₂ H ₆ + H \rightarrow C ₂ H ₅ + H ₂	/kC-H+H	/wCH3
n-butane \rightarrow 1-butene + H ₂	/kCC-HH	/wCH2; CH3
n-butane \rightarrow (<i>Z</i>)-2-butene + H ₂	/kCC-HH	/wCH2; CH2
CH ₃ CH ₃ \rightarrow CH ₃ + CH ₃	/kC-C	/wCH3; CH3
CH ₃ OH \rightarrow CH ₃ + OH	/kC-O	/wCH3; OH

Examples of Unimolecular Reaction Classes

Unimolecular: Radical Eliminations

- Bond Fission
- Beta Scission
- Cyclization

Unimolecular: Molecular Eliminations

- 1,2-Elimination
- 1,1-Elimination
- Cycloelimination
- Retroene-elimination

Unimolecular: Isomerizations

- 1,n Atom Transfer
- Bond Migration
- 1,5-sigmatropic Shift
- 3,3-Sigmatropic Rearrangement
- Bridged Isomerization
- Structural Isomerization
- Geometric Isomerization

Examples of Bimolecular/Composite Reaction Classes

Bimolecular

- Abstraction
- Radical Disproportionation
- Substitution
- Displacement

Composite

- Combination/Bond Fission
- Addition/Beta Scission

Examples of Radical Eliminations

Reaction Class	Reaction Form
<i>Bond Fission</i>	$R-X \rightarrow R\cdot + X\cdot$
<i>Beta Scission</i>	$\cdot RS-X \rightarrow R=S + X\cdot$
<i>Cyclization(*)</i>	$\cdot RS-X \rightarrow -(RS)- + X\cdot$

Reaction Class	Reaction	Connect	Rxn Class
<i>Bond Fission</i>	$CH_3CH_3 \rightarrow CH_3\cdot + CH_3\cdot$	/r1-2	/k C-C
<i>Beta Scission</i>	$CH_3CH_2O\cdot \rightarrow CH_2=O + CH_3\cdot$	/r3.2-1	/kO. C-C
<i>Cyclization</i> <i>(unimolecular substitution)</i>	$\cdot CH_2CH_2CH_2OOH \rightarrow -(CH_2CH_2CH_2O)- + \cdot OH$	/r1+5-4	/kC+O-O(4)

Examples of Molecular Eliminations

Reaction Class	Reaction Form
<i>1,2-Elimination</i>	$R(X)S(Y) \rightarrow R=S + XY$
<i>1,1-Elimination</i>	$R(X)(Y) \rightarrow R: + XY$
<i>Cycloelimination</i>	$R(X)S(Y) \rightarrow R=S + X=Y$
<i>Retroene-elimination</i>	$R(T=X)S(Y) \rightarrow R=S + T=XY$

Rxn Class	Reaction	Rxn Connect	Rxn Class
<i>1,2-Elim</i>	$C_2H_5OH \rightarrow CH_2=CH_2 + H_2O$	/r1,2- 3,4	/kCC- OH
<i>1,1-Elim</i>	$CH_3CH_3 \rightarrow CH_3CH: + H_2$	/r1- 3,4	/kC- HH
<i>Cycloelim</i>	Cyclohexene \rightarrow 1,3-Butadiene + C_2H_4	/r3*4- 5,6	/kCC- CC(4)
<i>Retroene</i>	1-Pentene \rightarrow Propene + C_2H_4	/r4,2- 1*5^6	/kCC- CC^H(4)

Examples of Isomerizations

Reaction Class	Reaction Form
1,n Atom Transfer	$\bullet\text{RS-X} \rightarrow \text{X-RS}\bullet$
Bond Migration	$\text{RS-X} \rightarrow \text{X-RS}$
1,5-Sigmatropic Shift	$\text{R=TS-X} \rightarrow \text{X-RT=S}$
3,3-Sigmatropic Rearrangement	$\text{R=TU=S} \rightarrow \text{T=RS=U}$
Bridged Isomerization	$\text{R(XY)S} \rightarrow \text{RS(X)(Y)}$
Structural Isomerization	$\text{X-R=S-Y} \rightarrow \text{Y-R=S-X}$
Geometric Isomerization	$\text{X/RS\Y} \rightarrow \text{X/RS/Y}$

Reaction Class	Reaction	Rxn Connect	Chem Sites
1,n Atom Transfer	$\bullet\text{CH}_2\text{CH}_2\text{OH} \rightarrow \text{CH}_3\text{CH}_2\text{O}\bullet$	/r3-4+1.	/kC.O^H(3)
Bond Migration	$\text{CH}_2=\text{CH}-\text{CH}_2\text{CH}_3 \rightarrow \text{CH}_3-\text{CH}=\text{CH}-\text{CH}_3$	/r3,4-5+1*3	/kCC^H(3)
1,5-Sigmatropic Shift	1,3-Hexadiene \rightarrow 2,4-Hexadiene	/r5,6-7+1*3	/kCC^H(5)
3,3-Sigmatropic	$\text{CT}_2=\text{CH}-\text{CD}_2\text{O}-\text{CH}=\text{CH}_2 \rightarrow \text{CD}_2=\text{CH}-\text{CT}_2\text{CH}_2-\text{CH}(=\text{O})$	/r1+2*3,5-6,4	/kCO^CC
Bridged Isomerization	$:\text{SiSiH}_2 \rightarrow \text{Si}(\text{H}_2)\text{Si}$	/r1-(3,4)+2	/kSiSi^HH
Struct Isomerization	Fulvene \rightarrow Benzene	/r6-1,4:6-1,7	/kC-CC:C-CH
Geom Isomerization	$(Z)\text{-CHCl=CHCl} \rightarrow (E)\text{-CHCl=CHCl}$	/r2~1	/kCC~ClCl

Examples of Bimolecular Reaction Classes

Reaction Class	Reaction Form
Abstraction	$R-X + Y\cdot \rightarrow R\cdot + XY$
Disproportionation	$\cdot RS-X + Y\cdot \rightarrow R=S + XY$
Substitution	$R-X + Y\cdot \rightarrow R-Y + X\cdot$
Displacement	$R-X + S-Y \rightarrow R-Y + S-X$

Reaction Class	Reaction	Rxn Connect	Rxn Class
Abstraction	$CH_3CH_3 + \cdot OH \rightarrow C_2H_5\cdot + H_2O$	/r2-3+1'	/kC-H+O
Disproportionation	$CH_3CH_2O\cdot + CH_3\cdot \rightarrow CH_3CH(=O) + CH_4$	/r1.2-4+1'	/kO.C-H+O
Substitution	$CH_3OOH + CH_3\cdot \rightarrow CH_3OCH_3 + \cdot OH$	/r3+1'-2	/kO+C-O
Displacement	$GeH_3SH + SiH_3OH \rightarrow GeH_3OH + SiH_3SH$	/r1,2'-2,1'	/kGe,O-S,Si

Examples of Composite Reactions

Reaction Class	Reaction Form
Combination/Bond Fission	$R\cdot + X\cdot \rightarrow S\cdot + Y\cdot$
Addition/Beta Scission	$R=S-Y + X \rightarrow R=S-X + Y$

Reaction Class	Reaction	Rxn Connect	Rxn Class
Combination /Bond Fission	$C_2H_5O\cdot + H\cdot \rightarrow$ $C_2H_5\cdot + \cdot OH$	/r1+1';1-2	/kC-H;C-O
Addition /Beta Scission	$CH_2=CH(OH) + CH_3\cdot \rightarrow$ $CH_2=CH-CH_3 + \cdot OH$	/r1.2+1';1.2-3	/kC.C-C;C.C-O

Summary of InChI-ER Layers

We propose five additional hierarchical layers for InChI (elementary reactions)

1. Reaction Formula [/v]. Identifies reactants and products, and their connectivities. The ordering of the reactants and products are “normalized” so that no matter how the reaction is “written,” it can be identified by a single and unique string.

2. Reaction Connectivity [/r]. Identifies bonds (connectivities) that are being broken and formed during the reaction.

3. Transition State Connectivity [/z]. Identifies the transition state for the reaction including only atoms actively involved in the reaction, and excluding spectator atoms.

4. Reaction Class [/k]. Characterizes the class (or type) of reaction. A symbolic string is used to identify different reaction types such as bond fission, abstraction, isomerizations, and 1,2-elimination. The form of the string identifies the general reaction class (e.g., R–X+Y for an abstraction), while the specific reaction class is identified by the atoms involved (e.g., C–H+H).

5. Chemical Sites [/w]. Identifies the chemical sites (groups) involved in the reaction. The hybridization of each site, the number of hydrogen atoms, and adjacent functionalities are provided.

InChI-ER

Apply InChI Methodology to Elementary Reactions

- Propose method of extending IUPAC International Chemical Identifier (InChI) to describe and identify elementary reactions.
- Based on existing InChI formalism and concepts, and fully extensible.
- Five layers proposed
 - **Reaction Formula** (A + B → C + D)
 - **Reaction Connectivity** (bonds broken and formed)
 - **Transition State Connectivity** (transition state structure)
 - **Reaction Class** (e.g. bond fission, abstraction)
 - **Chemical Sites** (chemical groups)
- Uniquely identifies elementary reactions (including chemical information).
- Provides chemical information to organize and discover reactions.
- Most layers determined from molecular structures for reactants/products.
- Applied to a representative large hydrocarbon combustion reaction set (based on USC Mech II and JetSurf 2.0).