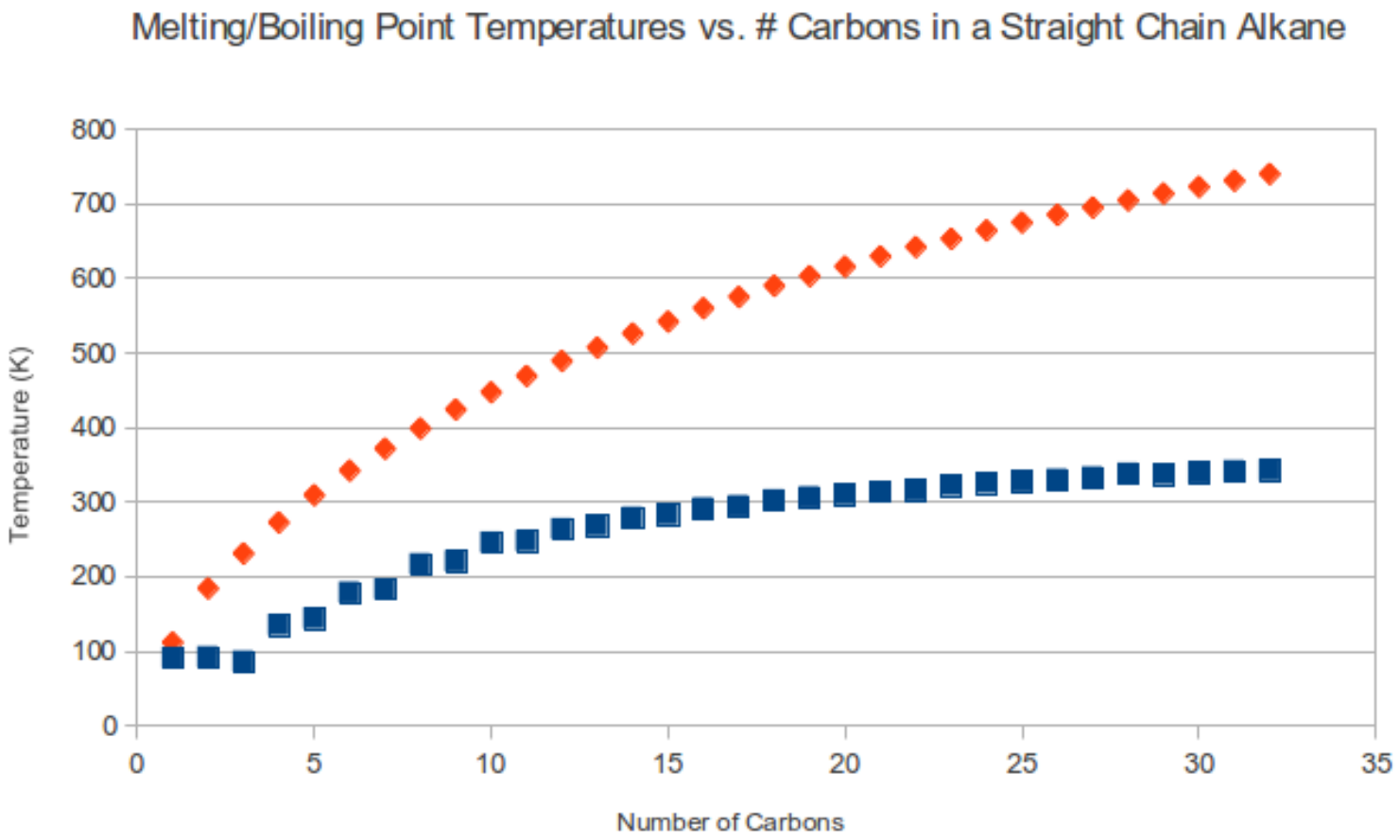
(Q)SAR Modelling

(Q)SAR = (Quantitative) structure-activity relationship

Regrese na strukturovaných datech


Melting/Boiling Point Temperatures vs. # Carbons in a Straight Chain Alkane
Temperature (K)
800
700
600
500
400
300
200
100
0
0
5
10
15
20
25
30
35
Number of Carbons

Klasický přístup

- Určíme vhodné atributy molekuly
  - Globální vlastnosti: počet atomů daného prvku, hydrofobicita, molekulární hmotnost…
  - Lokální vlastnosti: výskyty substruktur / jejich počet
- Preprocesujeme data
- Naučíme regresor / klasifikátor
  - Lineární regrese, rozhodovací strom, ANN, SVM...

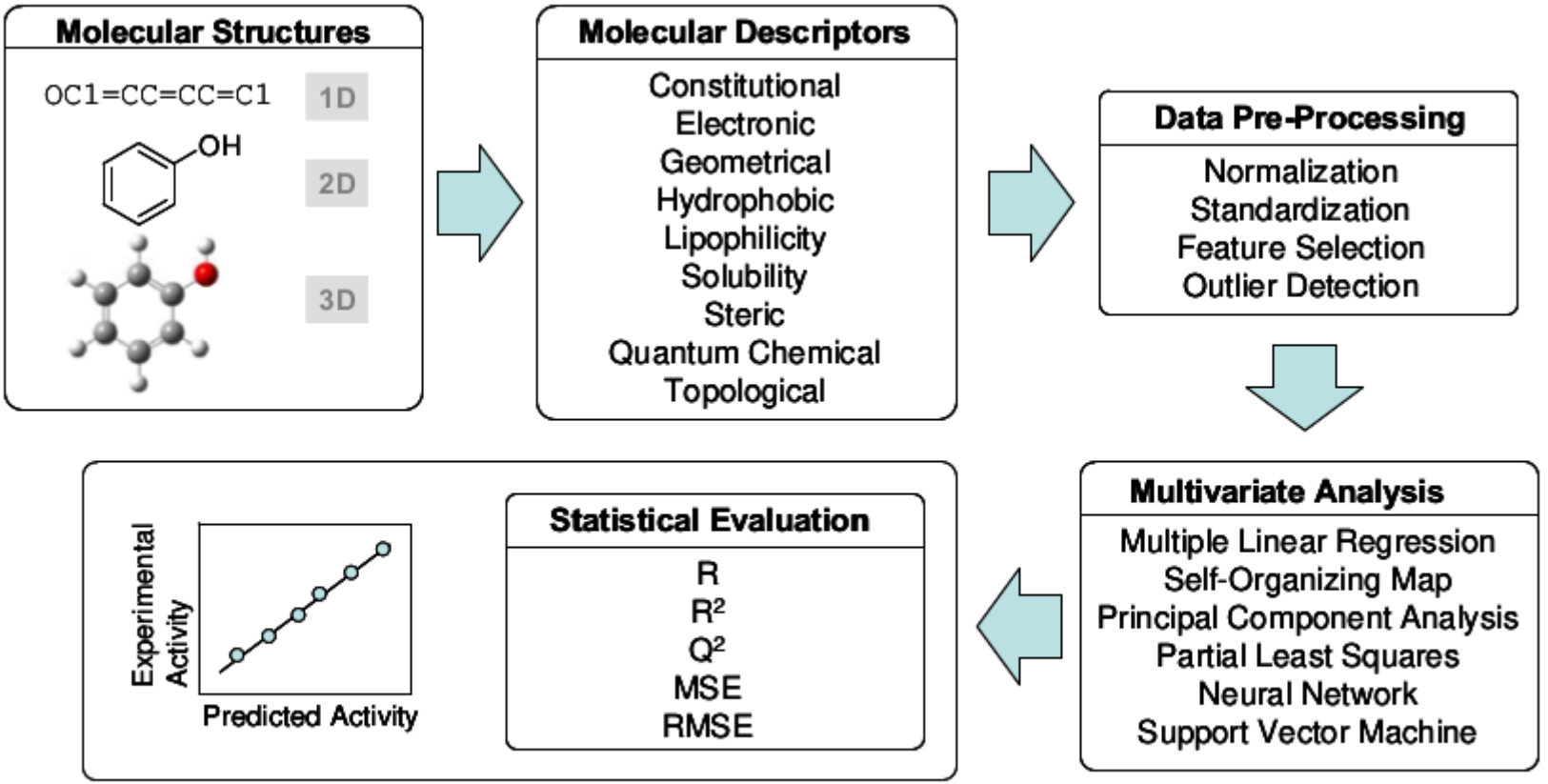


Figure 2: Schematic overview of the QSAR process.

Nevýhody

- Není jasné jak určit vhodné atributy
- Měření jejich hodnot mohou být nákladná
- Nutně ignorujeme velké množství strukturních informací
- Interpretace regresoru nemusí být v jazyce domény zřejmá

ILP Přístup

- Vstupní data zakódována do relační podoby
  - Můžeme zachovat kompletní strukturní informaci
- Výstupem je hypotéza o příčině studované vlastnosti vyjádřená logickou formulí

ILP – Formulace

- *B* – Logický program v Hornových klauzulích (Background knowledge), *E+* a *E-* pozitivní resp. negativní literály (pozorování). Hledáme hypotézu *h* takovou, že:

$$
\begin{array}{lll}
\textbf{Necessity:} & B & \not\models E^+ \\
\textbf{Sufficiency:} & B \wedge h & \models E^+ \\
\textbf{Weak consistency:} & B \wedge h & \not\models \mathit{false} \\
\textbf{Strong consistency:} & B \wedge h \wedge E^- & \not\models \mathit{false}
\end{array}
$$

Příklad reprezentace

Description	Representation	Comment
m1 is mutagenic	mutagenic(m1)	Facts, usually 100s
m1 has an atom a1	has_atom(m1,a1,carbon)	Facts, usually 1000s
a1 is at X,Y,Z in some conformation c1	has_atom(m1,a1,c1,-2.0,2.1,2.0)	Facts, usually 1000s
a1 and a2 are connected	has_bond(m1,a1,a2,double)	Facts, usually 1000s
m1 has a hydrogen donor	hdonor(m1,a10)	Rules for donors
m1 has a metal binding site at X,Y,Z in conformation c1	binds(m1,zinc,c1,0.8,0.3,1,2)	Stereochemistry rules
distance between a1 and a2 is 4 $\pm$ 1 A	dist(m1,a1,a2,4,0,1.0)	Geometry rules

Příklad naučené hypotézy

```
mutagenic(Chemical):-
        has_bond(Chemical,Atom1,Atom2,double),
        has_5_ring(Chemical,Ring),
        aromatic(Ring),
        in_ring(Atom1,Ring).
```

```
A compound is highly mutagenic if:
        a LUMO value $\leq$ -1.176 and an
        aryl-aryl bond between benzene rings

A chemical is highly mutagenic if:
        it contains a double bond conjugated to a
        5-membered aromatic ring via a carbon atom
```

Implementace

- PROGOL
- GOLEM
- DUCE
- CIGOL

Sloučení s klasickým přístupem

- Z naučených pravidel vytvoříme nové boolevské atributy
  - Pravidla můžeme generovat na podmnožinách dat
- Ověříme kvalitu nově nafitováne regresní funkce a signifikanci atributu

Propozicionalizace

- Jiný přístup k automatizovanému generování atributů
- Definice: Klauzule A, B jsou ekvivalentní když existuje substituce θ že: $A\theta \subseteq B$ a $B\theta \subseteq A$
- Definice: Klauzule je ireducibilní když je má nejnižší počet literálů v třídě ekvivalence.
- Algoritmus RelF: Generuje ireducibilní klauzule pokrývající (ve smyslu subsumpce) co nejvíce pozitivních a co nejméně negativních příkladů.
  - Formule mají tree-like vlastnost
  - Tvar formulí omezen zadanou šablonou
  - Umí počítat kolikrát se atribut v příkladu vyskytuje (tj. počet různých groundings atributu v příkladu)

Výhody použití logických atributů

- Snadná interpretovatelnost v jazyce domény
  - „Knowledge discovery“
- Větší využití strukturní informace v SAR
- Můžeme automaticky vygenerovat velké množství atributů
  - Ty by jinak musel sestavovat expert
- Lze kombinovat s globálními atributy při regresi / klasifikaci.