# Discrete Taylor Series as a Simple Way to Predict Properties of Chemical Substances like Benzenes and Cubanes

Jaime Nava and Vladik Kreinovich

Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
jenava@miners.utep.edu
vladik@utep.edu

November 5, 2009

# Introduction to the Practical Problem

- Many substances are obtained from a template molecule by replacing its H atoms with *ligands*.
- Examples of templates: benzene $C_6H_6$ and cubane $C_8H_8$.

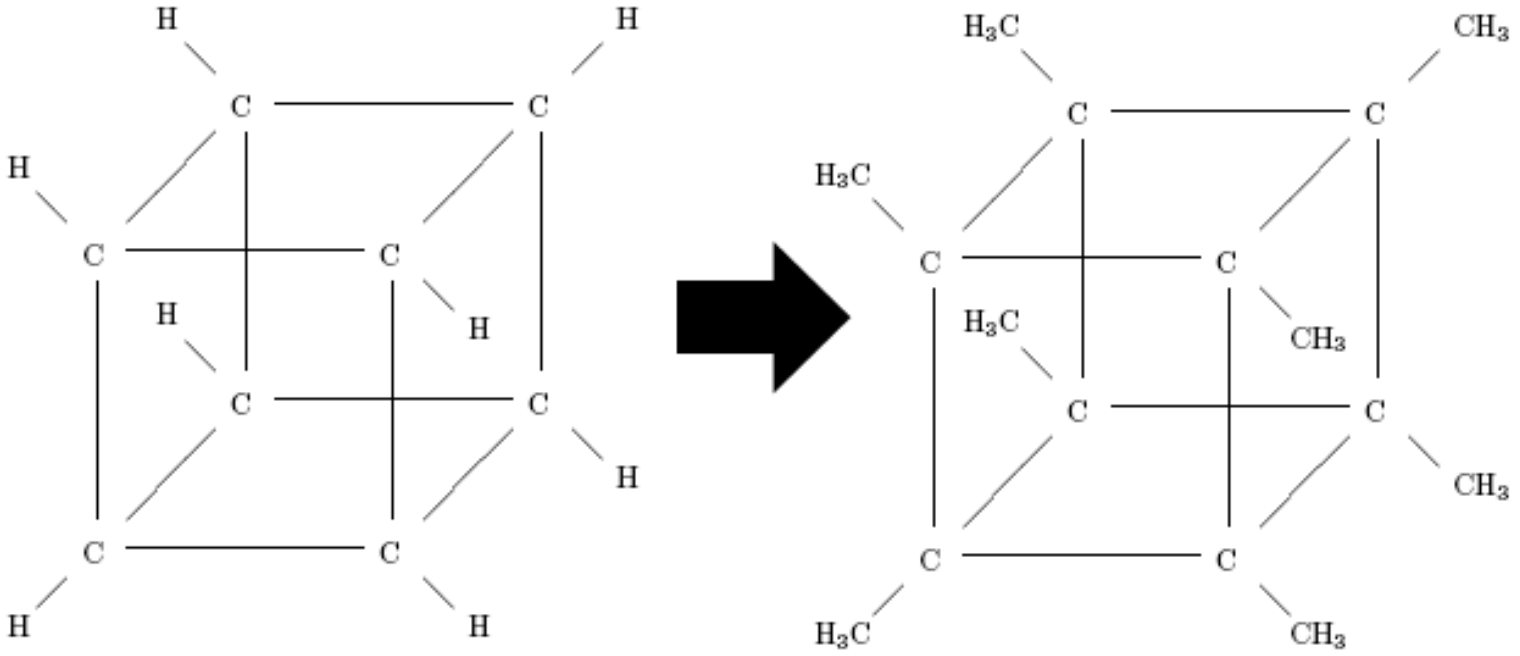



Figure: Example of a cubane derivative

# Introduction to the Practical Problem (cont-d)

- For each template, there are many possible substances.
- It is desirable to predict properties of these substances based on results of measuring a few of them.
- Such predictions are very important, since, e.g. cubanes, while kinetically stable, are highly explosive.
- As a result, at present, they are actively used as high-density, high-energy fuels and explosives.
- Researchers are investigating the potential of using cubanes in medicine and nanotechnology.

# How This Problem Is Solved Now and What We Propose

- It is desirable to predict properties of these substances based on results of measuring a few of them.
- Current approach proposed by D. J. Klein et al. in their 2007 Journal of Mathematical Chemistry.
- They used the ideas of the famous MIT mathematician Gian-Carlo Rota on partially ordered sets.
- Klein et al. showed that accurate predictions can be obtained by using these ideas.
- In this talk, we show that similar predictions can be made by using much simpler Taylor series techniques.

# Step-by-step Transitions as a Way to Predict Properties of Derivative Compounds

- Natural idea of synthesis: add ligands one by one
- Problem: use the properties of mono- and di-substituted substances to predict the properties of others

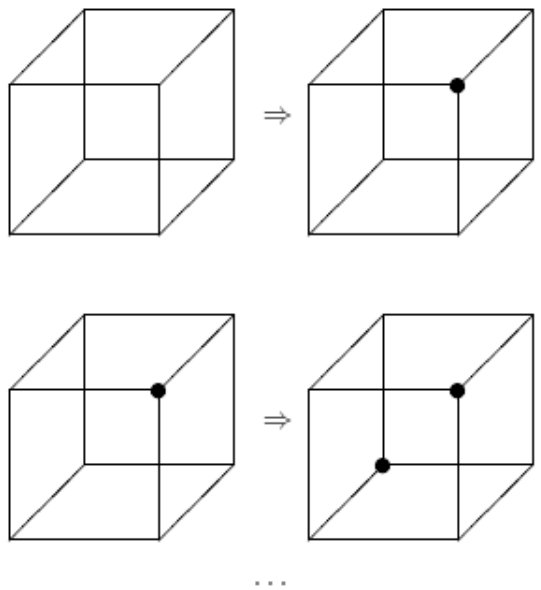


Figure: Step-by-step synthesis

# Another Practically Important Example: Benzene-Based Molecules

- Most organic molecules contain derivatives of benzene $C_6H_6$
- Adding ligands to benzene is one of the main ways to synthesize new organic molecules

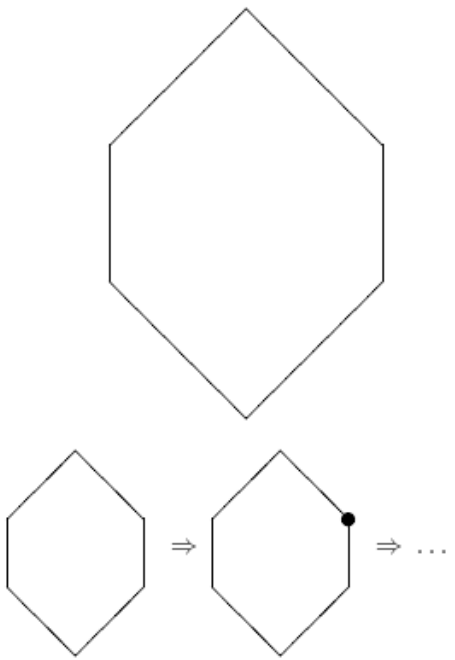



Figure: Benzene step-by-step substitution

# The Importance of Symmetry

- Molecules such as benzene or cubane have the additional property of symmetry
- Rotation does not change the chemical properties of a molecule
- This helps us reduce the number of tested substances

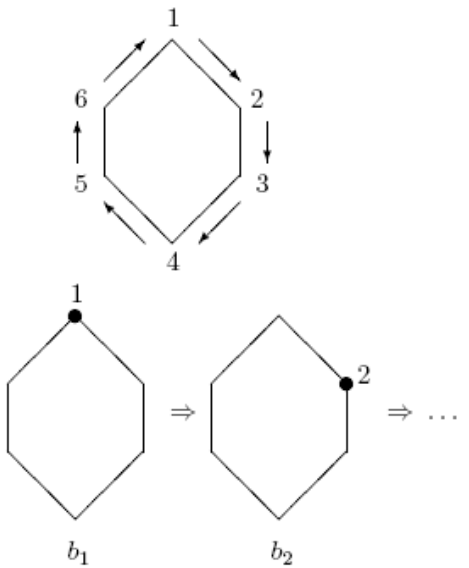



Figure: Benzene – rotation by 60°

# Extrapolation Example: Toxicity

- We perform measurements for several substances.
- We want to predict the values for all other substances.

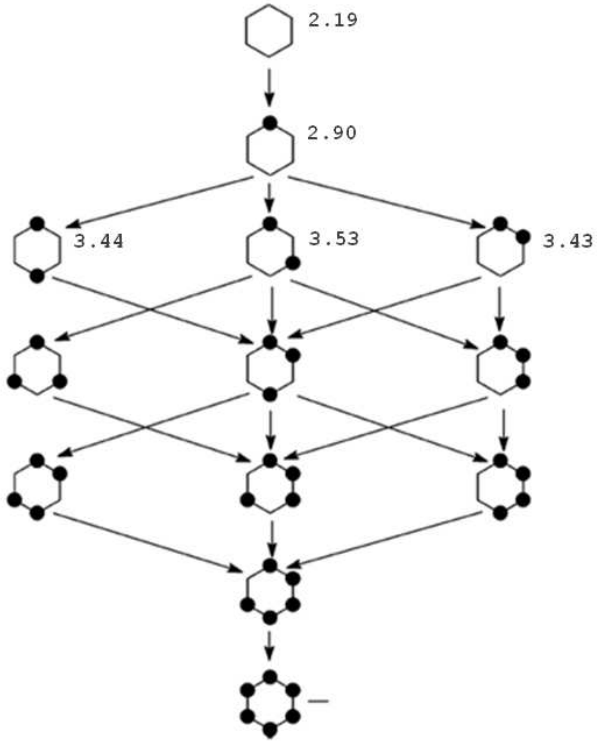



Figure: Extrapolation Example

# Heuristic Extrapolation Algorithms Based on Posets

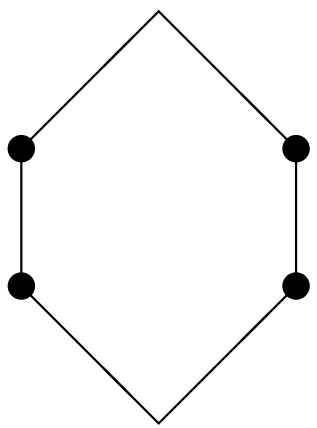
- In many practical situations, there is a natural partial order $x \le y$
- Example: $x \le y$ if a chemical substance $y$ can be obtained from a substance $x$ by a reaction of certain type
- Prediction on partially ordered sets (posets) according to Gian-Carlo Rota (MIT):
  - in general, we have $v(a) = \sum_{b:b \le a} V(b)$ for some $V(b)$
  - in practice, some values $V(b)$ are negligible, so we set $V(b) = 0$ for these $b$
- Resulting heuristic algorithm for estimating $v(a)$ from $v(a_1), \ldots, v(a_n)$:
  - use Least Squares method to find values $V(b)$ from the equations $v(a_i) \approx \sum_{b:b \le a_i} V(b)$, $1 \le i \le n$
  - use $\sum_{b:b \le a} V(b)$ as a predicted value of $v(a)$

# Poset-Based Extrapolation in Organic Chemistry

- General idea: $v(a) = \sum_{b:\, b \le a} V(b)$
- Natural relation $a \le b$: $b$ is obtained by $a$ by substituting some Hs by ligands
- Empirical fact (D. J. Klein et al.): $V(b) \approx 0$ for 3- and more-substituted molecules $b$
- Result: estimate $v(a)$ as $\sum_{b:\, b \le a} V(b)$, where $b$ goes over 0-, 1-, and 2-substituted molecules.
- For the original molecule $b$, we have $v(b) = V(b)$
- For a monosubstituted molecule $b_1$, we have $v(b_1) = V(b) + V(b_1)$, so $V(b_1) = v(b_1) - V(b)$
- For a disubstituted molecule $b_{12}$, $v(b_{12}) = V(b) + V(b_1) + V(b_2) + V(b_{12})$, so $V(b_{12}) = v(b_{12}) - V(b) - V(b_1) - V(b_2)$

# The Use of Symmetry

- Natural symmetries simplify the problem
- Example: all monosubstituted molecules are equivalent: $V(b_1) = V(b_2) = \ldots$
- Conclusion: $v(b_{12}) = V(b) + V(b_1) + V(b_2) + V(b_{12})$ becomes $v(b_{12}) = V(b) + 2V(b_1) + V(b_{12})$
- General case: $v(a) = \sum_{b:\, b \leq a} n(b) \cdot V(b)$. Example:



$$v(a) = V(b_0) + 4V(b_1) + 2V(b_{12}) + 2V(b_{13}) + 2V(b_{14}).$$

# Extrapolation Example: Toxicity

- Perform measurements for $v(a_1), \ldots, v(a_m)$.

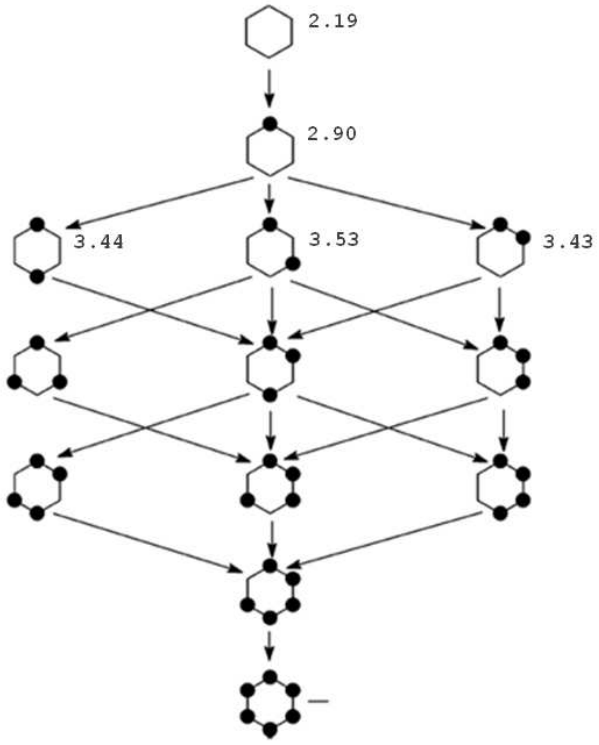



Figure: Extrapolation Example

# Extrapolation Example (cont-d)

- Use least squares to find values $V(b)$

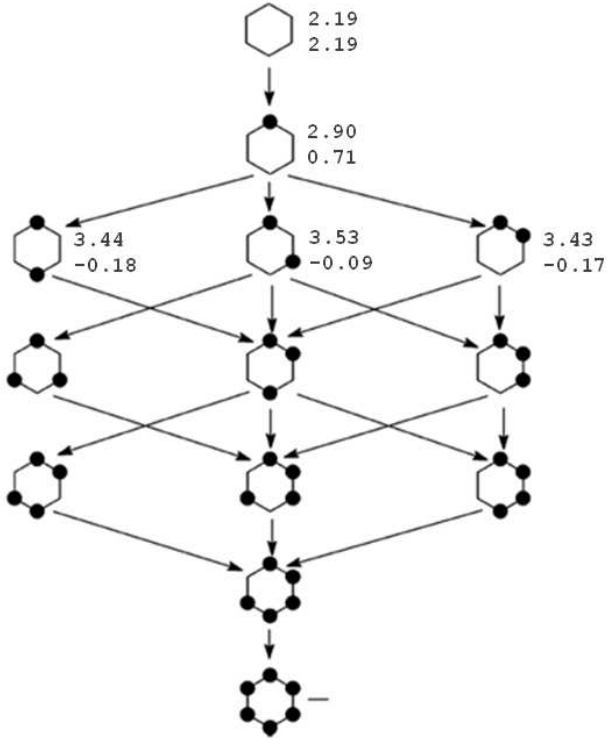



Figure: Extrapolation Example (cont-d)

# Extrapolation Example (cont-d)

- Extrapolate using the formula $v(a) = \sum_{b:b \leq a} V(b)$

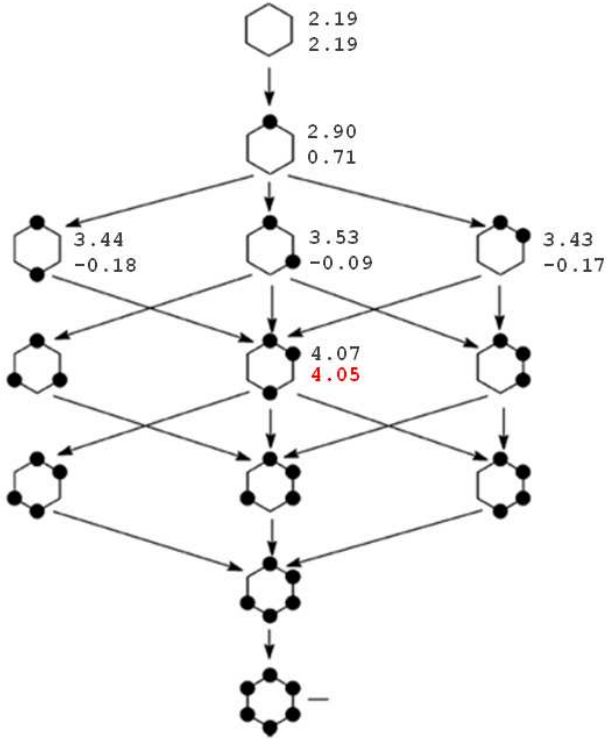



Figure: Poset Extrapolation

# Poset-Based Extrapolation is Empirically Successful

- Reminder: we extrapolate by using a heuristic formula

$$v(a) = \sum_{b:\, b \le a} n(b) \cdot V(b).$$

- Reminder: $n(b)$ is the number of $b$-type substances from which $a$ can be obtained by substitution.
- The resulting formulas lead to very good quality predictions of different quantities:
  - energy
  - boiling point
  - vapor pressure at a certain temperature
  - etc.

# Limitations of the Poset Approach and Our Work

- Poset-related approaches have relatively few empirically successful applications
- As a result, researchers may not have high confidence in these results
- To increase confidence, it is desirable to justify this heuristic approach
- We justify poset approach by using a technique with a much longer history of successful applications: Taylor series
- Specifically, we show that poset-based approach is equivalent to the approach based on Taylor series

# Taylor Series: A Standard Tool for Solving (Continuous) Problems in Science and Engineering

- In physical and engineering applications, most parameters $x_1, \ldots, x_n$ (coordinates, velocity, etc.) are *continuous*
- The dependence $y = f(x_1, \ldots, x_n)$ is also usually continuous and smooth (differentiable)
- Smooth functions can be usually expanded into Taylor series around some point $\widetilde{x} = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$:

$$f(x_1, \ldots, x_n) = f(\widetilde{x}_1, \ldots, \widetilde{x}_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \cdot \Delta x_i +$$

$$\frac{1}{2} \cdot \sum_{i=1}^{n} \sum_{i'=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_{i'}} \cdot \Delta x_i \cdot \Delta x_{i'} + \ldots,$$

where $\Delta x_i \stackrel{\text{def}}{=} x_i - \widetilde{x}_i$

# Taylor Series (cont-d)

- In practice, we can ignore higher-order terms
- Example: if linear approximation is not accurate enough, we can use quadratic approximation
- If we do not know the exact expression for $f(x_1, \ldots, x_n)$, we do not know the values of its derivatives $\dfrac{\partial f}{\partial x_i}$ and $\dfrac{\partial^2 f}{\partial x_i \partial x_j}$
- All we know is that we approximate a general function by a general linear or quadratic formula

$$f(x_1, \ldots, x_n) \approx c_0 + \sum_{i=1}^{n} c_i \cdot \Delta x_i + \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ii'} \cdot \Delta x_i \cdot \Delta x_{i'}$$

- The values of the coefficients $c_0$, $c_i$, and (if needed) $c_{ii'}$ can then be determined experimentally

# From Continuous to Discrete Taylor Series

- For each possible ligand location $i$, let $x_{i1}, \ldots, x_{ij}, \ldots, x_{iN}$ be parameters characterizing this location.
- Examples:
  - the density at a certain point,
  - the distance to a certain atom,
  - the angle between this atom and the given direction,
  - the angle describing the direction of the spin, etc.
- We are interested in the situations in which, at each location, there is either a ligand, or there is no ligand.
- For each location $i$ and for each parameter $x_{ij}$:
  - let $x_{ij}^-$ denote the value of the $j$-th parameter in the situation with no ligand at the location $i$, and
  - let $x_{ij}^+$ denote the value of the $j$-th parameter in the situation with a ligand at the location $i$.
- Default: when there is no ligand, i.e., $x_{ij} = x_{ij}^-$

# From Continuous to Discrete Taylor Series (cont-d)

- General case: $y = f(x_{11}, \ldots, x_{1N}, \ldots, x_{n1}, \ldots, x_{nN})$, so

$$y = y_0 + \sum_{i=1}^{n}\sum_{j=1}^{N} y_{ij} \cdot \Delta x_{ij} + \sum_{i=1}^{n}\sum_{j=1}^{N}\sum_{i'=1}^{n}\sum_{j'=1}^{N} y_{ij,i'j'} \cdot \Delta x_{ij} \cdot \Delta x_{i'j'},$$

  where $\Delta x_{ij} \stackrel{\text{def}}{=} x_{ij} - x_{ij}^{-}$
- Let $\varepsilon_i$ describe the presence of the ligand at the location $i$:
  - when there is no ligand, $\varepsilon_i = 0$, and
  - when there is a ligand, $\varepsilon_i = 1$.
- Reminder: $x_{ij} = x_{ij}^{-}$ if no ligand, $x_{ij} = x_{ij}^{+}$ if ligand
- General formula: $\Delta x_{ij} = \varepsilon_i \cdot \Delta_{ij}$, where $\Delta_{ij} \stackrel{\text{def}}{=} x_{ij}^{+} - x_{ij}^{-}$
- Substituting $\Delta x_{ij} = \varepsilon_i \cdot \Delta_{ij}$ into the formula for $y$:

$$y = a_0 + \sum_{i=1}^{n} a_i \cdot \varepsilon_i + \sum_{i=1}^{n}\sum_{i'=1}^{n} a_{ii'} \cdot \varepsilon_i \cdot \varepsilon_{i'}, \text{ where } a_i = \sum_{j=1}^{N} y_{ij} \cdot \Delta_{ij}$$

# Discrete Taylor Expansions can be Further Simplified

- First, for each $\varepsilon_i \in \{0, 1\}$, we have $\varepsilon_i^2 = \varepsilon_i$
- Thus, the term $a_{ii'} \cdot \varepsilon_i \cdot \varepsilon_{i'}$ corresponding to $i = i'$ is equal to $a_{ii} \cdot \varepsilon_i$, hence

$$y = c_0 + \sum_{i=1}^{n} c_i \cdot \varepsilon_i + \sum_{i \neq i'} c_{ii'} \cdot \varepsilon_i \cdot \varepsilon_{i'},$$

  where $c_0 = a_0$, $c_{ii'} = a_{ii'}$, and $c_i = a_i + a_{ii}$
- Second, we combine terms proportional to $\varepsilon_i \cdot \varepsilon_{i'}$ and to $\varepsilon_{i'} \cdot \varepsilon_i$
- As a result, we obtain the following simplified expression

$$y = v_0 + \sum_{i=1}^{n} v_i \cdot \varepsilon_i + \sum_{i < i'} v_{ii'} \cdot \varepsilon_i \cdot \varepsilon_{i'},$$

  where $v_0 = c_0$ and $v_{ii'} = c_{ii'} + c_{i'i}$

# Example

$$y = a_0 + \sum_{i=1}^{n} a_i \cdot \varepsilon_i + \sum_{i<i'} a_{ii'} \cdot \varepsilon_i \cdot \varepsilon_{i'}$$


$$y = a_0$$


$$y = a_0 + a_1$$


$$y = a_0 + (a_1 + a_2) + a_{12}$$


$$y = a_0 + (a_1 + a_2 + a_4) + (a_{12} + a_{24} + a_{14})$$

# Comparing Poset and Discrete Taylor Series Approaches

- Reminder: $\varepsilon_i = 0$ means no ligand, $\varepsilon_i = 1$ means ligand
- Taylor series: $y = v_0 + \sum_{i=1}^{n} v_i \cdot \varepsilon_i + \sum_{i<i'} v_{ii'} \cdot \varepsilon_i \cdot \varepsilon_{i'}$
- Poset approach: $v(a) = \sum_{b:\, b \le a} V(b)$
- Here, $b \le a$ means that $a$ can be obtained from $b$ by adding ligands
- So, if $b = (\varepsilon'_1, \ldots, \varepsilon'_n)$ and $a = (\varepsilon_1, \ldots, \varepsilon_n)$, then $b \le a$ means that for every $i$, we have $\varepsilon'_i \le \varepsilon_i$
- Resulting formula:

$$y = V(a_0) + \sum_{i:\, \varepsilon_i = 1} V(a_i) + \sum_{i<i':\, \varepsilon_i = \varepsilon_{i'} = 1} V(a_{i,i'})$$

# Proof that The Discrete Taylor Series are Indeed Equivalent to the Poset Formula

- Taylor series: $y = v_0 + \sum_{i=1}^{n} v_i \cdot \varepsilon_i + \sum_{i<i'} v_{ii'} \cdot \varepsilon_i \cdot \varepsilon_{i'}$
- Poset: $y = V(a_0) + \sum_{i:\, \varepsilon_i = 1} V(a_i) + \sum_{i<i':\, \varepsilon_i = \varepsilon_{i'} = 1} V(a_{i,i'})$
- Proof that these formulas coincide:

$$\sum_{i:\, \varepsilon_i = 1} V(a_i) = \sum_{i:\, \varepsilon_i = 1} V(a_i) \cdot \varepsilon_i = \sum_{i=1}^{n} V(a_i) \cdot \varepsilon_i$$

$$\sum_{i<i':\, \varepsilon_i = \varepsilon_{i'} = 1} V(a_{i,i'}) = \sum_{i<i':\, \varepsilon_i = \varepsilon_{i'} = 1} V(a_{i,i'}) \cdot \varepsilon_i \cdot \varepsilon_j$$

$$\sum_{i<i':\, \varepsilon_i = \varepsilon_{i'} = 1} V(a_{i,i'}) = \sum_{i<i'} V(a_{i,i'}) \cdot \varepsilon_i \cdot \varepsilon_j$$

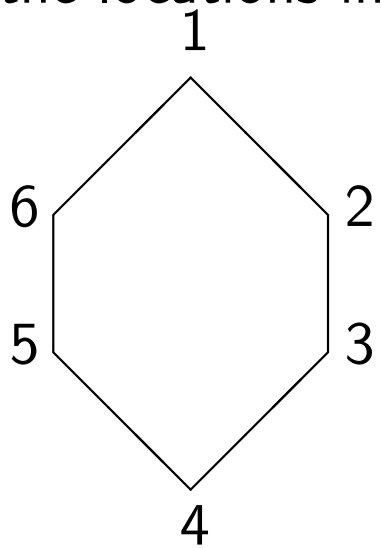
- So, we can take $v_0 = V(a_0)$, $v_i = V(a_i)$, $v_{ii'} = V(a_{i,i'})$

# Important Observation: The Presence of Symmetry Does Not Change the Equivalence

- Reminder: symmetry means that some of the coefficients $v_i$ and $v_{ii'}$ coincide.
- Example: for benzenes and cubanes, symmetry means
  - that $v_1 = v_2 = \ldots = v_i = \ldots$, and
  - that the value $v_{ii'}$ depends only on the distance $d(i, i')$ between the locations $i$ and $i'$
- Notations: $V \stackrel{\text{def}}{=} v_i$, $V_d$ denotes $v_{ii'}$ when $d(i, i') = d$

# Symmetry: Example

- Let us number the locations in a sequential order:



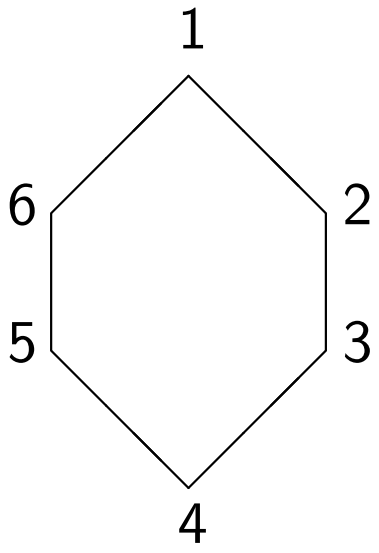

- In these notations, the general quadratic formula takes the form $y = v_0 + V \cdot \left(\sum_{i=1}^{n} \varepsilon_i\right) +$

$$V_1 \cdot (\varepsilon_1 \cdot \varepsilon_2 + \varepsilon_2 \cdot \varepsilon_3 + \varepsilon_3 \cdot \varepsilon_4 + \varepsilon_4 \cdot \varepsilon_5 + \varepsilon_5 \cdot \varepsilon_6 + \varepsilon_6 \cdot \varepsilon_1) +$$

$$V_2 \cdot (\varepsilon_1 \cdot \varepsilon_3 + \varepsilon_2 \cdot \varepsilon_4 + \varepsilon_3 \cdot \varepsilon_5 + \varepsilon_4 \cdot \varepsilon_6 + \varepsilon_5 \cdot \varepsilon_1 + \varepsilon_6 \cdot \varepsilon_2) +$$

$$V_3 \cdot (\varepsilon_1 \cdot \varepsilon_4 + \varepsilon_2 \cdot \varepsilon_5 + \varepsilon_3 \cdot \varepsilon_6)$$

# Symmetry: Example (cont-d)




In other words, we have $y = v_0 + V \cdot N + \sum_{d=1}^{3} V_d \cdot N_d$, where

- $N$ is the total number of ligands, and
- $N_d$ is the total number of pairs $(i, i')$ of ligands with $d(i, i') = d$

# Main Advantage of the Taylor Representation

- Taylor series is a more familiar technique for a wide range of scientists
- Taylor series have a much larger number of successful applications than the poset-related methods;
- Therefore, scientists are more confident in Taylor series techniques.

# Additional Advantage: Taylor Series can Clarify the Equivalence of Different Arrangements

- Consider, in the poset formulation, instead of the original order $b \le a$, the *dual* order $b \le' a$ which is defined as $a \le b$
- $a \le b$ means: we can obtain $b$ from $a$ by *adding* ligands
- The dual order $b \le' a$ means: we can obtain $b$ from $a$ by *removing* ligands
- In the original order $\le$, the minimal element is the original substance $a_0$
- 2nd order poset approximation means: use values $V(b)$ corresponding to substances with 0, 1, and 2 ligands
- In the dual order $\le'$, the minimal element is the substance with the ligands in all the places
- 2nd order poset approximation means: We use values $V(b)$ corresponding to substances with 0, 1, and 2 missing ligands

# Additional Advantage of the Taylor Representation (cont-d)

- Will this new order lead to a different approximation?
  - difficult to immediately answer this question
  - the two orders are different, it may look like the resulting approximations are different too.
- Reformulate this question in terms of the discrete Taylor series
  - $\varepsilon'_i = 0$ (no change) if there is a ligand at the $i$-th location
  - $\varepsilon'_i = 1$ (change) if there is no ligand at the $i$-th location
  - hence $\varepsilon'_i = 1 - \varepsilon_i$ and $\varepsilon_i = 1 - \varepsilon'_i$
  - so, expressions quadratic in $\varepsilon_i$ are also quadratic in $\varepsilon'_i$ and vice versa
  - conclusion: the resulting approximation is exactly the same for the new order

# Additional Advantage: A Detailed Description

- Reminder: $y = v_0 + \sum_{i=1}^{n} v_i \cdot \varepsilon_i + \sum_{i<i'} v_{ii'} \cdot \varepsilon_i \cdot \varepsilon_{i'}$
- We substitute $\varepsilon_i = 1 - \varepsilon'_i$ into this formula
- We get general quadratic formula

$$y = v_0 + \sum_{i=1}^{n} v_i \cdot (1 - \varepsilon'_i) + \sum_{i<i'} v_{ii'} \cdot (1 - \varepsilon_i) \cdot (1 - \varepsilon_{i'})$$

- Opening parentheses, we conclude that

$$y = v'_0 + \sum_{i=1}^{n} v'_i \cdot \varepsilon'_i + \sum_{i<i'} v'_{ii'} \cdot \varepsilon'_i \cdot \varepsilon'_{i'}, \text{ where}$$

$$v'_0 = v_0 + \sum_{i=1}^{n} v_i + \sum_{i<i'} v_{ii'}, \quad v'_i = -v_i - \sum_{i' : i<i'} v_{ii'} - \sum_{i' : i'<i} v_{i'i},$$

and $v'_{ii'} = v_{ii'}$

# Additional Advantage: A Detailed Description (cont-d)

- Similarly, we can start with

$$y = v_0' + \sum_{i=1}^{n} v_i' \cdot \varepsilon_i' + \sum_{i<i'} v_{ii'}' \cdot \varepsilon_i' \cdot \varepsilon_{i'}'$$

- We substitute $\varepsilon_i' = 1 - \varepsilon_i$
- We get

$$y = v_0 + \sum_{i=1}^{n} v_i \cdot \varepsilon_i + \sum_{i<i'} v_{ii'} \cdot \varepsilon_i \cdot \varepsilon_{i'},$$

where

$$v_0 = v_0' + \sum_{i=1}^{n} v_i' + \sum_{i<i'} v_{ii'}', \quad v_i = -v_i' - \sum_{i' : i<i'} v_{ii'}' - \sum_{i' : i'<i} v_{i'i}',$$

and $v_{ii'} = v_{ii'}'$.

# Example

- Reminder: for benzene, $y = v_0 + V \cdot N + \sum_{d=1}^{3} V_d \cdot N_d$, where
  - $N$ is the total number of ligands, and
  - $N_d$ is the total number of pairs $(i, i')$ of ligands with $d(i, i') = d$
- Here,

$$v_0' = v_0 + 6V + 6V_1 + 6V_2 + 3V_3$$
$$V' = -V - 2V_1 - 2V_2 - V_3$$
$$V_1' = V_1, \quad V_2' = V_2, \quad V_3' = V_3$$

- Correspondingly,

$$v_0 = v_0' + 6V' + 6V_1' + 6V_2' + 3V_3'$$
$$V = -V' - 2V_1' - 2V_2' - V_3'$$
$$V_1 = V_1', \quad V_2 = V_2', \quad V_3 = V_3'$$

# Conclusion

- Case study: predicting properties of new chemical substances
- Several chemical substances can be obtained by adding ligands to a "template" molecule like benzene $C_6H_6$ or cubane $C_8H_8$
- There is a large number of such substances, and it is difficult to synthesize all of them
- It is desirable to synthesize only a few of them and to extrapolate
- Such an extrapolation has been obtained by using Rota's ideas related to partially ordered sets
- We show that the same extrapolation follows from a more familiar mathematical technique: Taylor series
- This makes the chemical prediction results more reliable

# Acknowledgments


- We wish to extend our gratitude to Dr. Larry Ellzey for his encouragment
- Thanks to Dr. Guillermo Restrepo for his collaboration in the Chemistry case study