Mapping transition metal chemical space for machine learning models

Jon Paul Janet $^1$ Heather Kulik $^1$

$^1$Department of Chemical Engineering, Massachusetts Institute of Technology

2017 AIChE Annual Meeting
Minneapolis, MN

11.01.17
Data-driven molecular design

Machine learning is transforming how we design new materials...
Data-driven molecular design


Machine learning is transforming how we design new materials...
Data-driven molecular design

Machine learning is transforming how we design new materials...


OLED chemical space
Data-driven molecular design

Machine learning is transforming how we design new materials...


OLED chemical space

$\text{NN} \sim 10^6$
Data-driven molecular design


Machine learning is transforming how we design new materials...

OLED chemical space

NN $\sim 10^6$

DFT $\sim 10^5$
Data-driven molecular design

**Machine learning is transforming how we design new materials...**


OLED chemical space

- **NN $\sim 10^6$**
- **DFT $\sim 10^5$**
- **Exp. $\sim 10^1$**
Machine learning is transforming how we design new materials...

Machine learning is transforming how we design new materials...

...what about inorganic molecular complexes?

Data-driven molecular design

Machine learning is transforming how we design new materials...

...what about inorganic molecular complexes?

Data-driven molecular design

Machine learning is transforming how we design new materials...

...what about inorganic molecular complexes?

Data-driven molecular design

Machine learning is transforming how we design new materials...

...what about inorganic molecular complexes?


Data-driven molecular design

Machine learning is transforming how we design new materials...

...what about inorganic molecular complexes?


ML for TM complexes

Standard tools do not work well for TM complexes$^1$:  

\[ \text{test RMSE, kcal/mol} \]

\[ \begin{array}{cccc}
CM-ES & MCDL & \end{array} \]

ML for TM complexes

Standard tools do not work well for TM complexes:\n
1345 (194) complexes

7 HF values

---

Standard tools do not work well for TM complexes:\(^1\):
ML for TM complexes

Standard tools do not work well for TM complexes\textsuperscript{1}:

- 7 HF values
- 1345 (194) complexes

Metal ID: Fe

Oxidation: 2+

ML for TM complexes

Standard tools do not work well for TM complexes\(^1\):

- Metal ID: Fe
  - Oxidation: 2+

- Proximal properties
  - Max \(\Delta \chi\) and \(\sum \Delta \chi\) over all ligands
    - \(\chi = 3.04\)
    - \(M \rightarrow N \rightarrow C \rightarrow S\)
    - \(\chi = 2.55\)
    - \(\Delta \chi = 0.49\)

- Metal properties

---

ML for TM complexes

Standard tools do not work well for TM complexes\(^1\):

- **7 HF values**
- **1345 (194) complexes**

Metal ID: Fe

Oxidation: 2+

Metal properties

Max \( \Delta \chi \) and \( \sum \Delta \chi \) over all ligands

\[ \chi = 3.04 \]

\[ \chi = 2.55 \]

\[ \Rightarrow \Delta \chi = 0.49 \]

Proximal properties

Kier Index: \( \kappa_2 \)

Ligand properties

---

ML for TM complexes

Standard tools do not work well for TM complexes\(^1\):

![Graph showing test RMSE for CM-ES and MCDL methods]

ML for TM complexes

Standard tools do not work well for TM complexes\textsuperscript{1}:

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure.png}
\end{figure}

\begin{itemize}
\item 7 HF values
\item 1345 (194) complexes
\end{itemize}

\begin{itemize}
\end{itemize}
RAC descriptors

RACs based on autocorrelations\(^2\)

\[ \sum_{O} O Z O Z = 48 \]

\[ d_{1} = 48 + \sum_{C} C Z O Z C = 144 + 48 \]

\[ \sum_{i} \sum_{j} Z_i Z_j \delta(d_{ij}, 1) \]

\[ d_{x} = \sum_{i} \sum_{j} Z_i Z_j \delta(d_{ij}, x) \]

How to adapt to TM complexes?

restrict the scope to focus on near-metal atoms

\[ \sum_{M} O Z M Z \]

\[ d_{2} = \sum_{M} C Z M Z \]

\[ d_{3} = \sum_{M} O Z M Z O (Z_i - Z_j) \]

properties:

\[ T, \chi, Z, I, S \sim 160 \text{ features in total} \]

---

RAC descriptors

RACs based on autocorrelations\(^2\)

RAC descriptors

RACs based on autocorrelations\(^2\)

[Diagram of a molecular structure with labels for atoms and bonds]

RAC descriptors

RACs based on autocorrelations

\[ d_1 : \sum_{O,C} Z_O Z_C = 48 \]

\[ d_1 : \sum_{M} Z_M Z_O = 48 + 144 = 192 \]

\[ d_2 : \sum_{M,C} Z_M Z_C = 144 + 48 = 192 \]

\[ d_3 : \sum_{M} (Z_i - Z_j) \approx 160 \text{ features in total} \]

---

RAC descriptors

RACs based on autocorrelations\(^2\)

\[ d_1 : 48 + \sum_{C,O} Z_O Z_C = 144 + 48 \]

RAC descriptors

RACs based on autocorrelations\(^2\)

\[
d_1 : \sum_i \sum_j Z_i Z_j \delta(d_{i,j}, 1)
\]

---

RAC descriptors

RACs based on autocorrelations

\[ d_x : \sum_i \sum_j Z_iZ_j \delta(d_{ij}, x) \]

RAC descriptors

RACs based on autocorrelations\(^2\)

How to adapt to TM complexes?

RAC descriptors

RACs based on autocorrelations\(^2\)

\[ \sum_{M, O} Z_M Z_O = 48 \]
\[ \sum_{M, C} Z_M Z_C = 144 + 48 \]
\[ \sum_{i} \sum_{j} Z_i Z_j \delta(d_{ij}, 1) \]
\[ \sum_{i} \sum_{j} Z_i Z_j \delta(d_{ij}, x) \]

How to adapt to TM complexes?
restrict the scope to focus on
near-metal atoms

RAC descriptors

RACs based on autocorrelations

How to adapt to TM complexes?
restrict the scope to focus on
near-metal atoms

---

\[ \sum_{M, O} Z_M Z_O = 48 \]
\[ \sum_{M, C} Z_M Z_C = 144 + 48 \]
\[ \sum_{M, O} \left( Z_i - Z_j \right) \]

properties: \( T, \chi, Z, I, S \) \( \sim 160 \) features in total

---

RAC descriptors

RACs based on autocorrelations

How to adapt to TM complexes?
restrict the scope to focus on near-metal atoms

\[ d_1 : \sum_{M,O} Z_M Z_O \]

---

RAC descriptors

RACs based on autocorrelations

How to adapt to TM complexes?
restrict the scope to focus on
near-metal atoms

\[ d_2 : \sum_{M,C} Z_M Z_C \]

---

RAC descriptors

RACs based on autocorrelations\(^2\)

\[
\begin{align*}
    d_1 & : \sum M, O \\
    d_2 & : \sum M, C \\
    d_3 & : \sum_{M, O} Z_M Z_O \quad (Z_i - Z_j)
\end{align*}
\]

How to adapt to TM complexes? restrict the scope to focus on *near-metal atoms*

\(\text{properties: } T, \chi, Z, I, S \sim 160 \text{ features in total}\)

RAC descriptors

RACs based on autocorrelations\(^2\)

\[ d_1 : \sum_{i} \sum_{j} Z_i Z_j \delta(d_{ij}, 1) \]

\[ d_2 : \sum_{i} \sum_{j} Z_i Z_j \]

\[ d_3 : \sum_{M, O} Z_M Z_O \]

How to adapt to TM complexes?
restrict the scope to focus on
*near-metal atoms*

RAC descriptors

RACs based on autocorrelations\textsuperscript{2}

\[
d_1 : \sum M, O Z_M Z_O
\]

\[
d_2 : \sum M, C Z_M Z_C
\]

\[
d_3 : \sum_{M,O} \left( Z_i - Z_j \right)
\]

How to adapt to TM complexes?
restrict the scope to focus on
near-metal atoms

RAC descriptors

RACs based on autocorrelations

How to adapt to TM complexes? restrict the scope to focus on near-metal atoms

\[ d_3 : \sum_{M,O} Z_M Z_O (Z_i - Z_j) \]

properties: \( T, \chi, Z, I, S \)

---

RAC descriptors

RACs based on autocorrelations\(^2\)

How to adapt to TM complexes?
restrict the scope to focus on
near-metal atoms

\[
d_3 : \sum_{M,O} Z_M Z_O (Z_i - Z_j)
\]

\(~ 160\) features in total

Feature selection

Feature selection

Feature selection

Feature selection

Feature selection

Feature selection

Feature selection

Feature selection

Feature selection

Feature selection

Feature selection

Feature selection

Feature selection

Feature selection

Feature selection

RMSE, kcal/mol

MCDL
RAC155
UV86
RFE43
LS28
rF41

1.5
2.0
2.5
3.0
3.5
4.0

50 100 150
dimension

How local is too local?

How local is too local?

How local is too local?

How local is too local?

How local is too local?

Do features depend on properties?
Do features depend on properties?
Do features depend on properties?

All atoms equal
Do features depend on properties?

spin splitting (randF)
Do features depend on properties?

spin splitting (randF)

All atoms equal bond lengths (randF) redox (randF) more 'electronic'
more 'topological'
Do features depend on properties?

spin splitting (randF)

spin splitting (randF)
Do features depend on properties?

spin splitting (randF)

more ‘electronic’
Do features depend on properties?

spin splitting (randF)

bond lengths (randF)

spin splitting (randF)

more ‘electronic’
Do features depend on properties?

- spin splitting (randF)
- bond lengths (randF)
- redox (randF)

More 'electronic' vs. more 'topological'
Do features depend on properties?

- spin splitting (randF)
- bond lengths (randF)
- redox (randF)

More 'electronic'

More 'topological'

80%

40%
Mapping TM complex space

PC 1
PC 2
E^0 (eV)
3
5
7
9
11
Mapping TM complex space
Mapping TM complex space
Mapping TM complex space

PC 1

PC 2

E 0 (eV)
Mapping TM complex space
Mapping TM complex space
Mapping TM complex space
Acknowledgments

Conclusions:

- machine learning TM complexes faces unique challenges
- ACs are a promising starting point for low-cost descriptors
- different target properties depend on different physical variables and we can gain design insights but figuring out which