Creating an atomic species representation to improve machine learning models in the chemical sciences

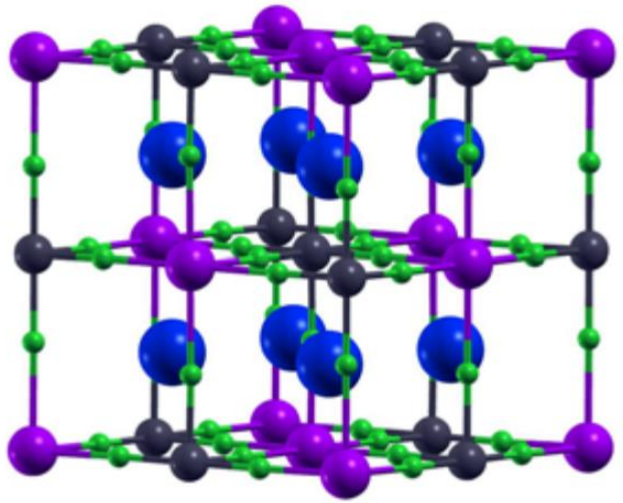
JOHN E. HERR, KEVIN KOH, KUN YAO, AND JOHN PARKHILL

Objetivo

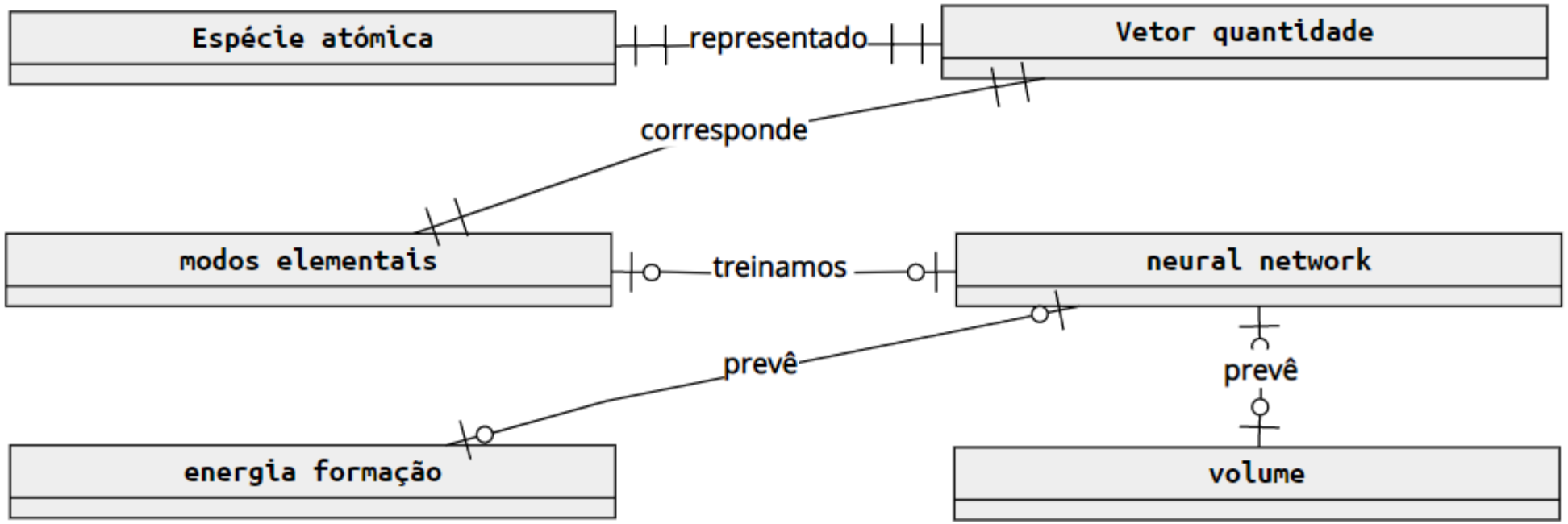
Melhorar os modelos de *machine learning* em ciências químicas para melhor representação de espécies atómicas (ex: elpasolites…)

Representação de estados químicos intermédios com vetores de estado para cálculo de energia livre em quebras/formação de ligações

ABC_2D_6



Metodologia

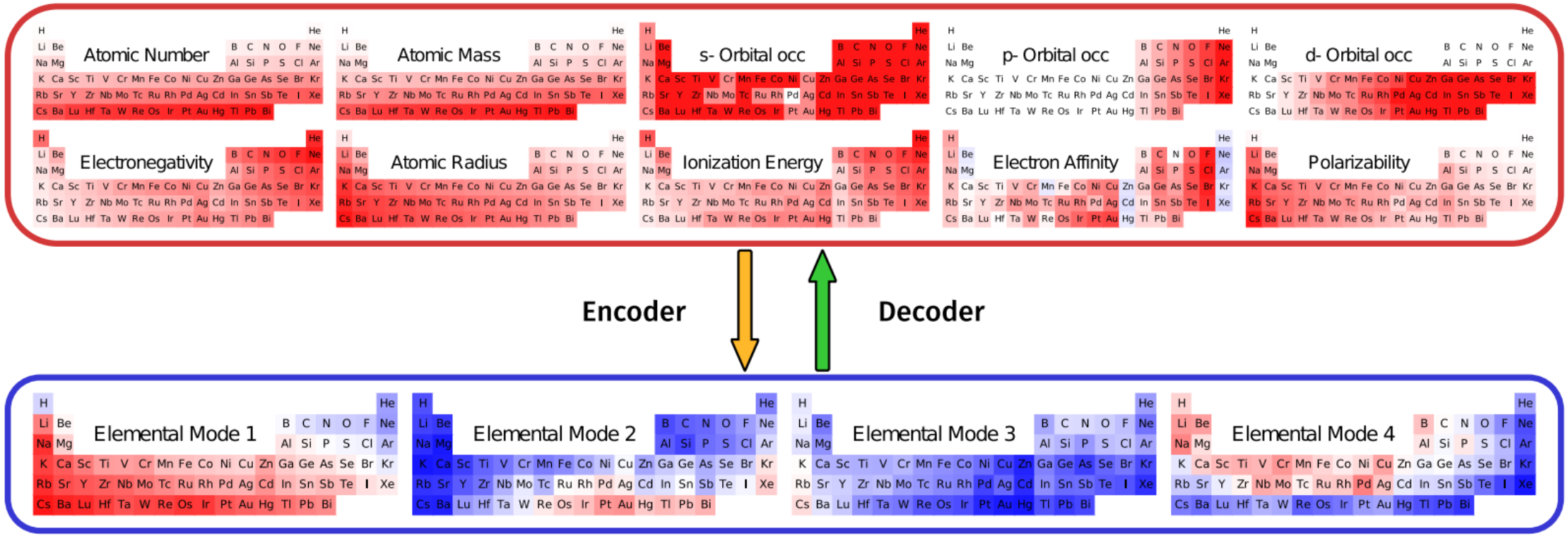


Scalling (Escalabilidade) e Eficiência

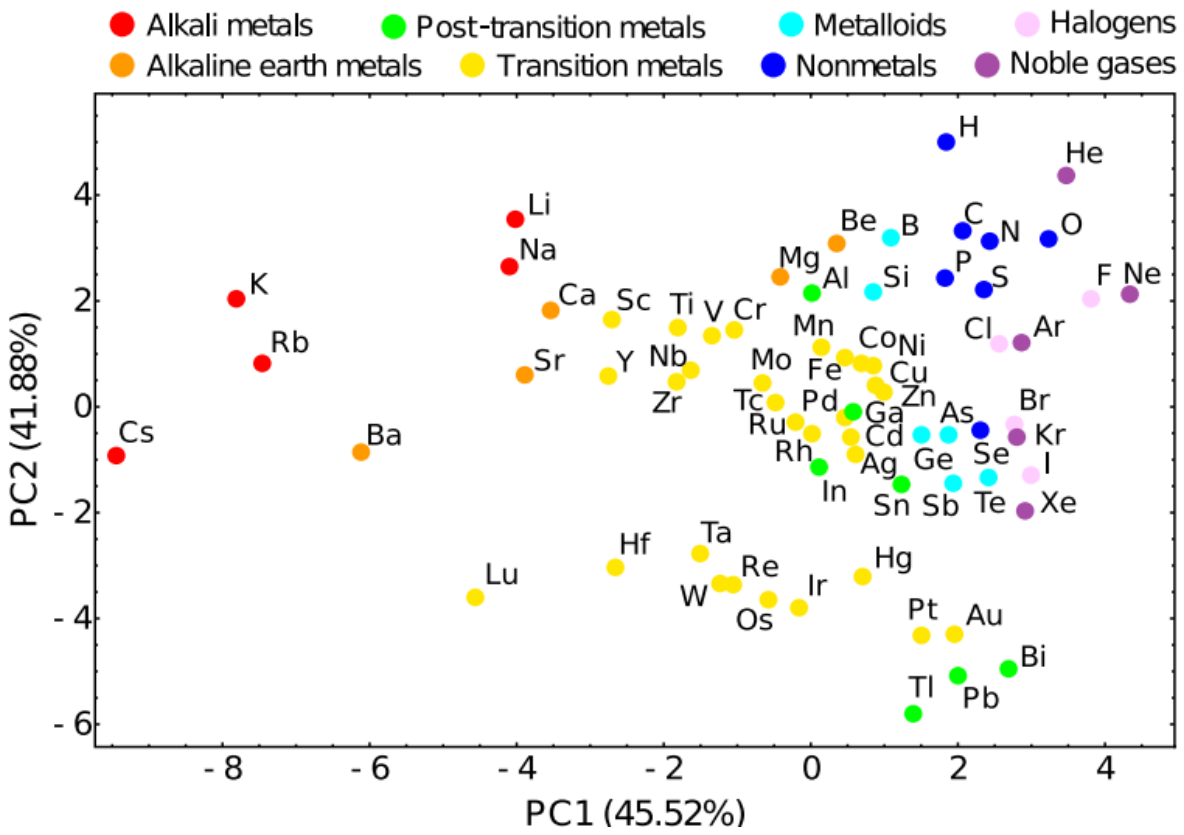
- Propriedade de um sistema de aguentar um aumento de trabalho adicionando recursos ao sistema
- Maior problema = maior uso de RAM
- Maior escalabilidade de certo problema = aumento de trabalho sem adicionar recursos

- Como se define vetores de *features* há melhor escalabilidade
- Passamos de um modelo com 4 elementos [H,C,N e O] para 11 elementos [H ,C, N, O, F, P, S, Cl, Se, Br e I]

Modos Elementais



Modos Elementais

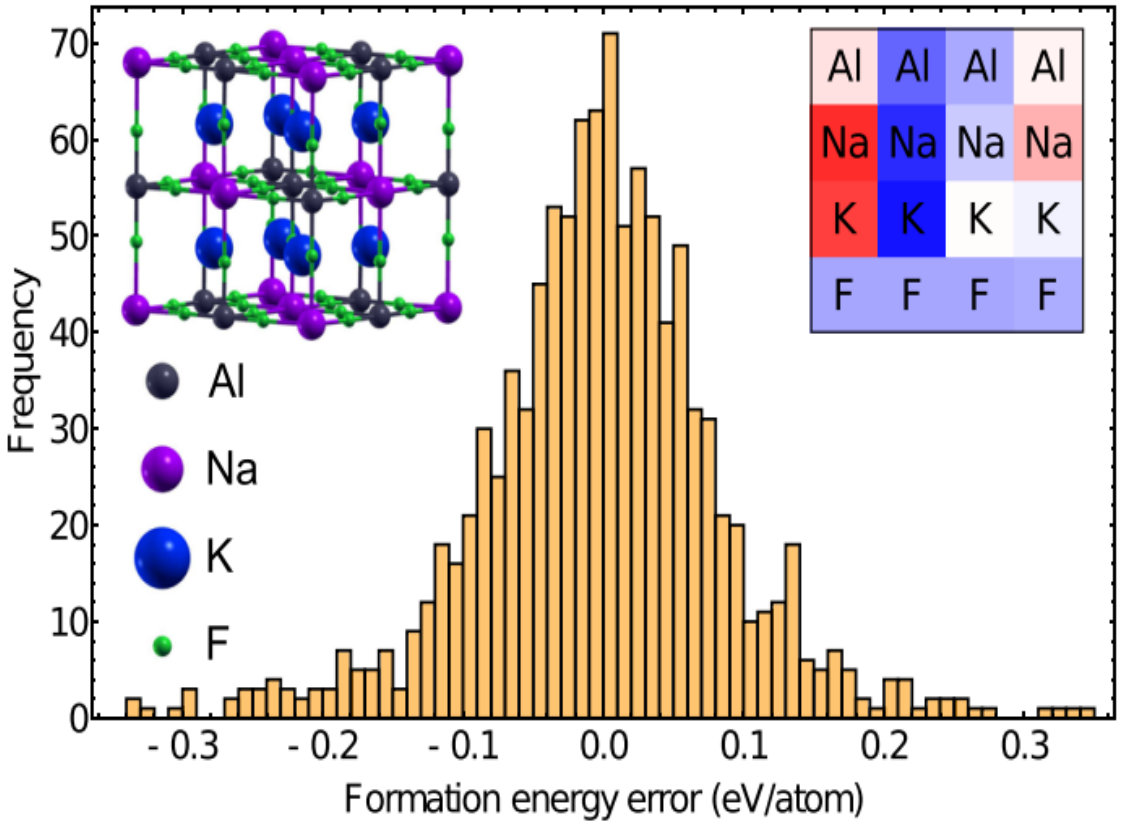


Variância *feature* com maior importância (PC1)

Variância 2nd *feature* maior importância (PC2)

Os modos elementais obtiveram informação pertinente para a diferenciação de espécies atómicas

Predição da Energia de Formação



O *dataset* é ≈ 10 000 cálculos de energia de formação na configuração ABC_2D_6.

O vetor de *features* é composto por 4 modos elementais correspondentes a cada elemento

A matriz vai ser colapsada segundo a direção do cristal antes de entrar na *neural network*

O *dataset* é separado em 80:10:10 (treino, teste e validação)

O MAE (mean absolute error) foi de apenas 67 meV/átomo, melhor do que qualquer modelo anterior

Modelo *neural network* químico

Ao contrário de cristais com estruturas bem definidas, moléculas orgânicas (C,H,N e O) apresentam um desafio muito maior, com 4 variáveis confinadas ao raio e 10 aos ângulos entre átomos

O vetor de *features* tem agora de fornecer informações geométricas do átomo e espécies envolventes

Modelo *neural network* químico

Uma forma de contornar os problemas de *scalling* e otimizar o modelo para certos sistemas, é manter o vetor de *features* com tamanho constante

Vamos aplicar modos elementais como fatores de acoplamento das variáveis das *atom-centered symmetry functions* (ACSFs)

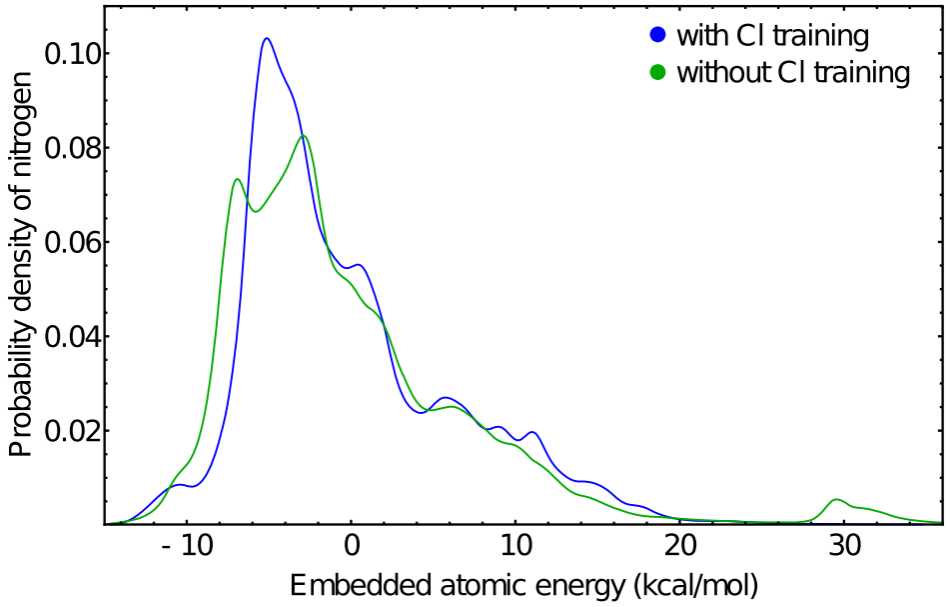
$$G^R_{\epsilon,R_s} = e^{-\eta\left(R_{ij}-R_s\right)^2} f_c(R_{ij}) \otimes \beta_\epsilon(Z_j)$$

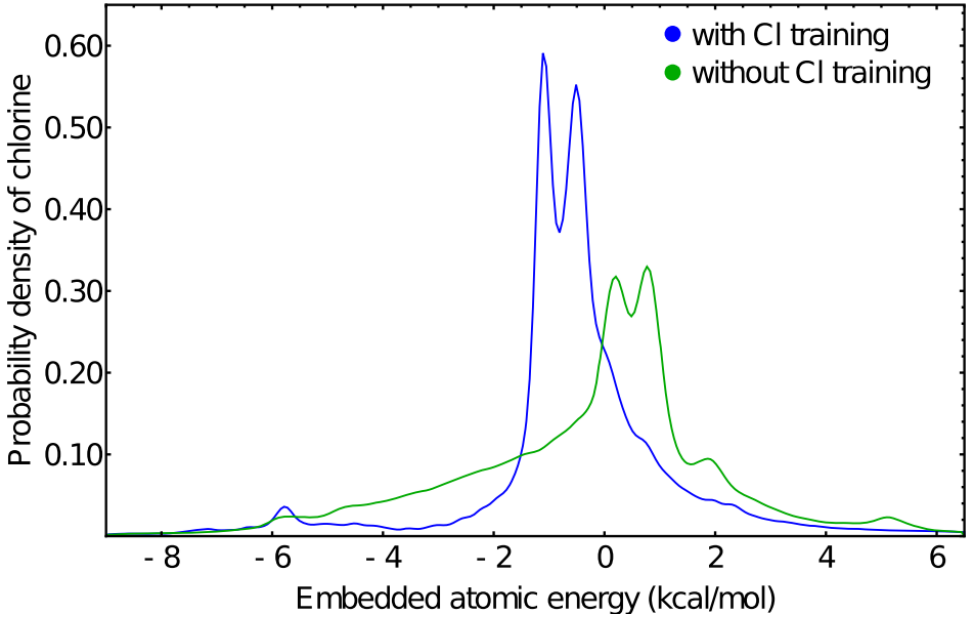
$$G^A_{\epsilon,R_s,\theta_s} = 2^{1-\zeta}\left(1+\cos\left(\theta_{ijk}-\theta_s\right)\right)^\zeta e^{-\eta\left(\frac{R_{ij}+R_{ik}}{2}-R_s\right)^2} \times f_c(R_{ij}) f_c(R_{ik}) \otimes \beta_\epsilon(Z_j)\beta_\epsilon(Z_k),$$

Resultados

Com um *dataset* com 65 000 moléculas únicas, treinamos o modelo e para verificar o quão bem o modelo relacionou informação

Clonar o *dataset* removendo moléculas Cl-, e voltar a treinar para ambos os casos





Resultados

A forma das distribuições é em grande parte igual

Para o *dataset* completo temos

- MAE: 1,31 kcal/mol
- RMSE: 1,63 kcal/mol
- MSD: -1,12 kcal/mol

Para o *dataset* sem moléculas com átomos Cl temos:

- MAE: 9,20 kcal/mol
- RMSE: 12,74 kcal/mol
- MSD: -6,43 kcal/mol

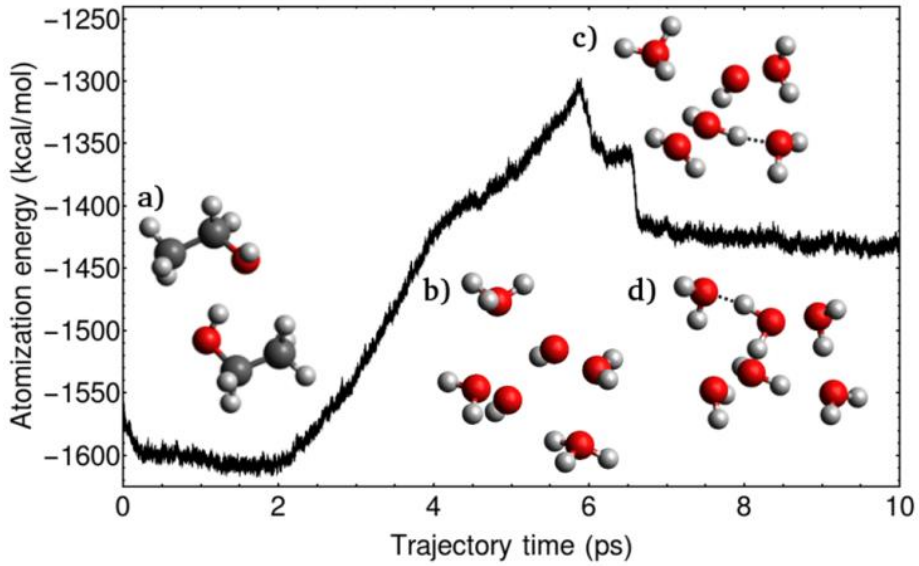
Provamos que a *neural network* transfere conhecimento, ajudando a generalizar para novas espécies atómicas com menos dados

Transformações químicas

A nossa *neural network* pode também avaliar os estados intermédios duma transformação interpolando o vetor de *features* entre o estado inicial e final

$$I_{i-j,\epsilon-\epsilon'} = (1-\lambda)I_{i,\epsilon} + \lambda I_{j,\epsilon'}$$

Para testar esta implementação vamos testar a transformação de dímeros de etanol em hexâmeros de água



Conclusões

A representação de espécies atómica é prevista comprimindo propriedades físicas inerentes a cada partícula em espaços dimensionais pequenos através de um *autoencoder*

Os modelos elementares provaram se ótimos em descrever materiais assim como os *elpasolitos*, assim percebendo qual as estruturas mais estáveis para pesquisa experimental

Os modelos elementares são também bons a parametrizar *neural networks* de modelos químicos