

I.A. « LA REVOLUTION GNoME », par Dr Kheireddin KADRI / AptisSkills

The speaker



Kheireddin Kadri/ Researcher Data scientist



ESILV
ENGINEERING SCHOOL
DE VINCI PARIS



Société de consultants –
Pôle R&D AS+ESILV (composants hydrogène) =>



1 - AU COMMENCEMENT

Les gnomes selon la kabbale,

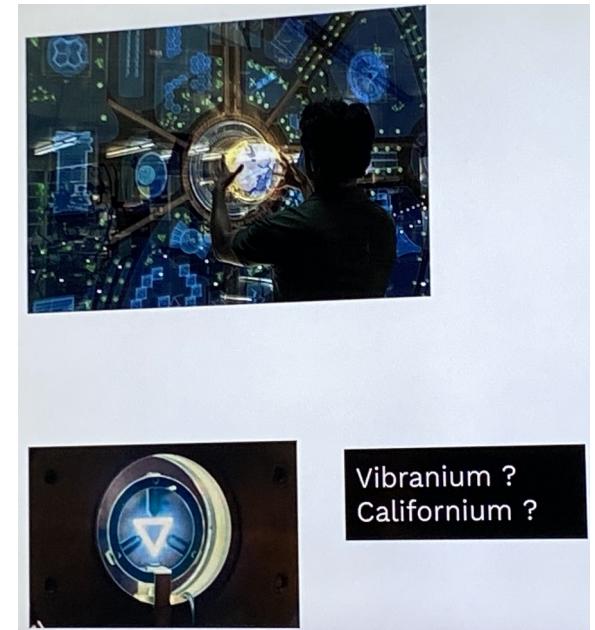
De la référence à l'univers de J.R.R. Tolkien, où le Mithril, matériau fait la destinée de l'Univers... à celui de Stan LEE (Ironman), avec le Vibranium et le Californium, ces matériaux conduisent à faire un pont, via la donnée (data), vers le Big Data (le « sexe des ados).



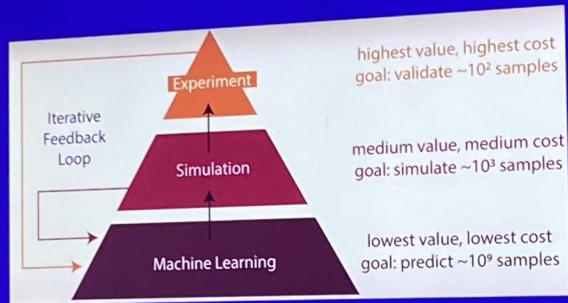
Est-ce une bénédiction ou une malédiction ?

=> Un constat :

L'augmentation de productions d'articles scientifiques sur les nouveaux composants de batterie conduit à l'impossibilité de suivre la recherche.
Nécessité de recourir aux nouvelles technologies.



Cursed or Blessed

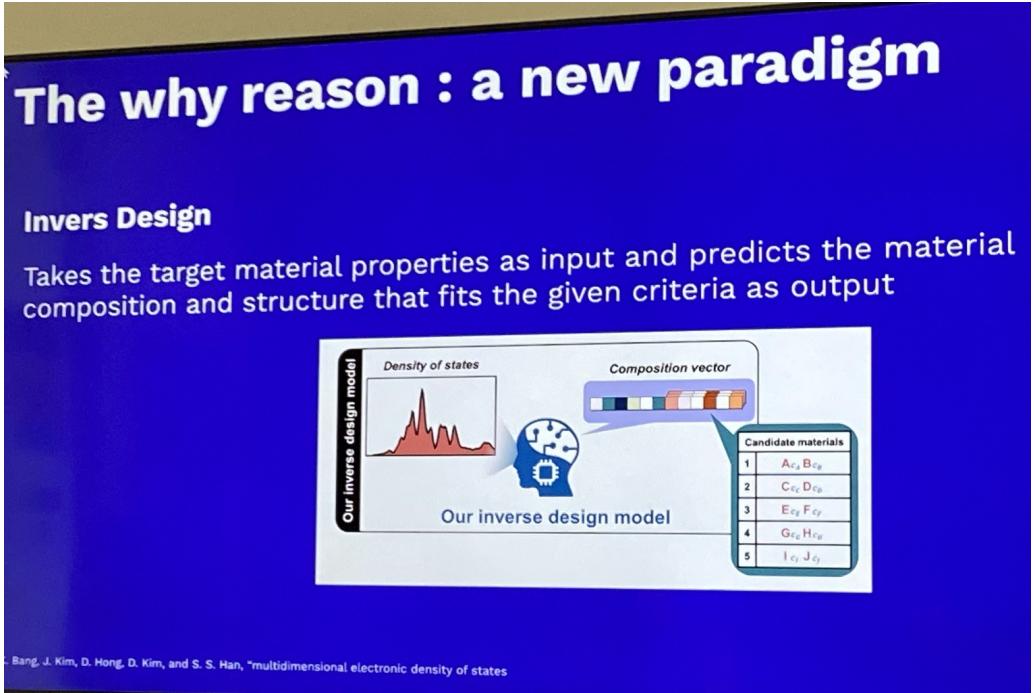


Materials Informatics
Taylor Sparks : <https://www.youtube.com/@TaylorSparks>

Cercle vertueux (processus itératif) entre synthèse et abondance de la donnée.

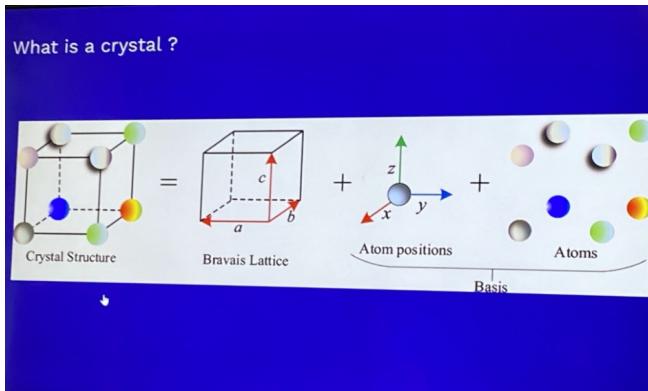
Pourquoi ? ... Pour faire émerger un nouveau paradigme

=> Prédire un matériel en faisant un salto arrière grâce à un input (densité d'état grâce à une short list).
Le saut est difficile et comporte des limites...

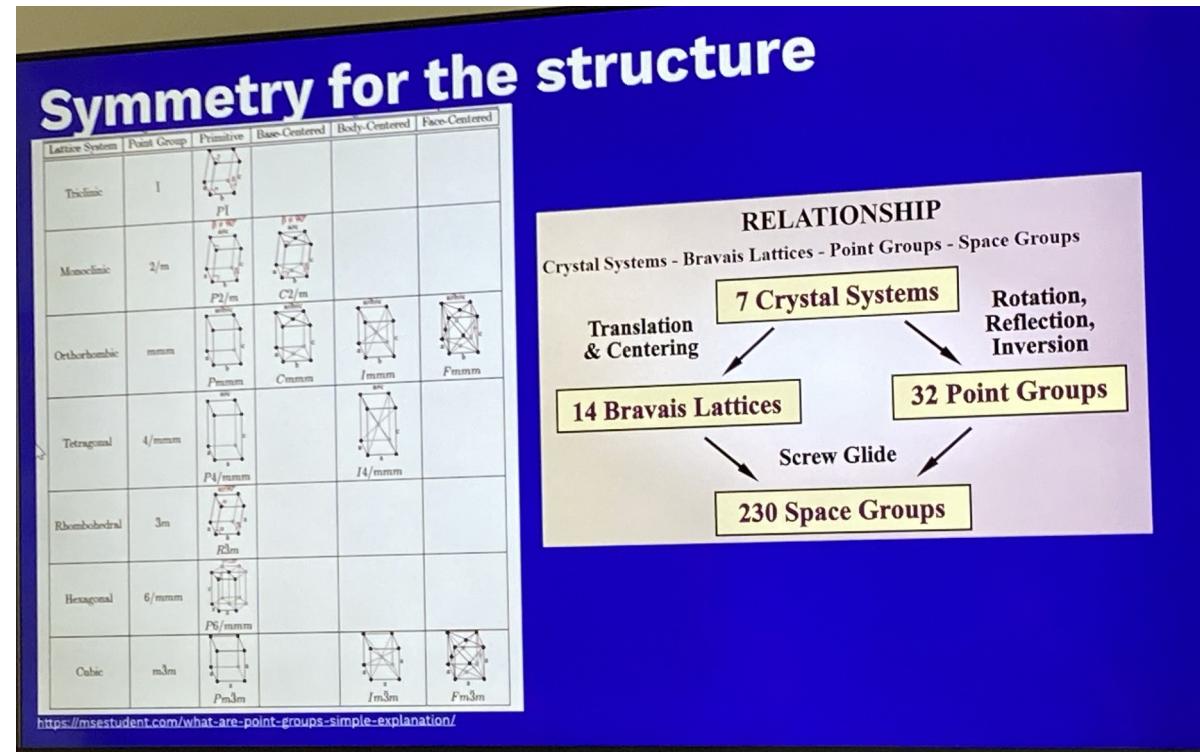
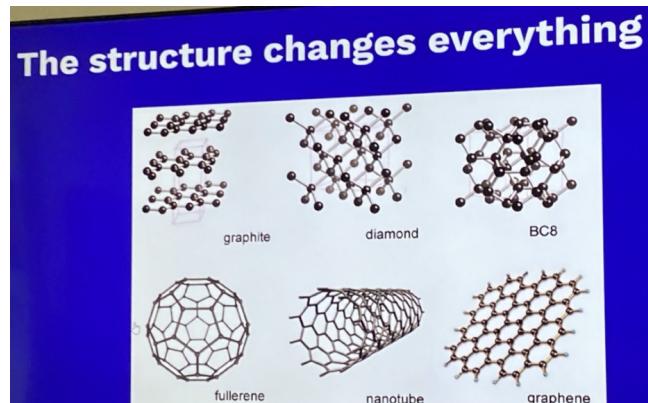


Comment ?

... A partir de la recette du cristal :
Étudier place des atomes et leurs compositions.

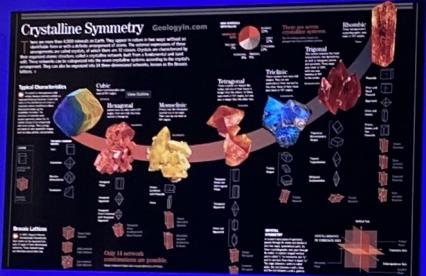


Exemple du carbone => On change la structure :



=> Le seul fait d'ajouter une symétrie par translation permet d'obtenir 14 formes ; puis, avec rotation, on obtient 32 ; puis, par translation avec torsion on arrive à 230 groupes d'espaces qui sont agnostiques (structure en dehors de l'atome).

Law of crystal symmetry

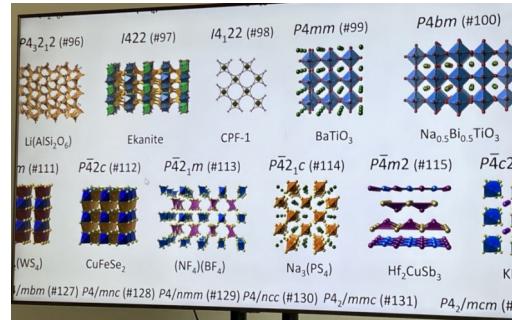


https://www.xtal.iqf.csic.es/Cristalografia/parte_03_1-en.html

| Graphic symbol | Num. symbol | Graphic symbol | Num. symbol |
|----------------|-------------------|----------------|-------------------|
| None | 1 | ○ | 1 |
| 1 | 2 | ◊ | 2/m |
| 2 ₁ | 2 ₁ /m | ◊ | 2 ₁ /m |
| 3 | 3 | △ | 3 |
| 3 ₁ | 3 ₁ | ◊ | 4 |
| 3 ₂ | 3 ₂ | ◊ | 4/m |
| 4 | 4 | ◊ | 4 ₁ /m |
| 4 ₁ | 4 ₁ | ◊ | 6 |
| 4 ₂ | 4 ₂ | ◊ | 6/m |
| 4 ₃ | 4 ₃ | ◊ | 6 ₃ /m |
| 6 | 6 | m | m |
| 6 ₁ | 6 ₁ | a, b, c | a, b, c |
| 6 ₂ | 6 ₂ | n | n |
| 6 ₃ | 6 ₃ | d | d |
| 6 ₄ | 6 ₄ | 1/2 | 1/2 |
| 6 ₅ | 6 ₅ | 3/2 | 3/2 |
| 6 ₆ | 6 ₆ | d | d |

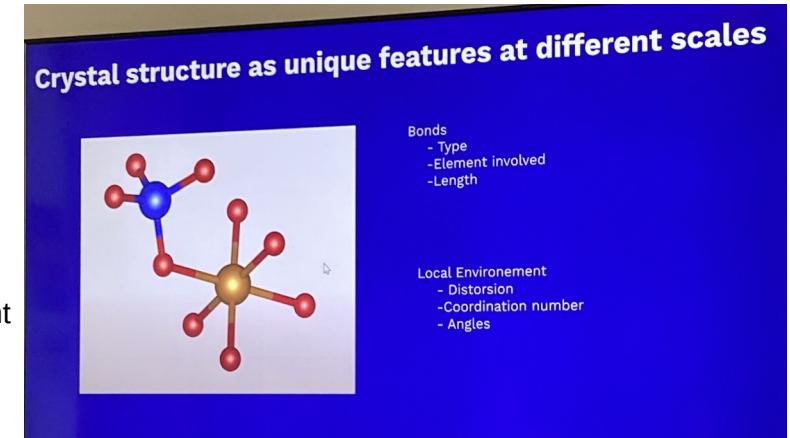
Utiliser pour les composants photovoltaïques : la symétrie miroir

=> les outils actuels permettent la visualisation multidimensionnelle de ces cristaux.



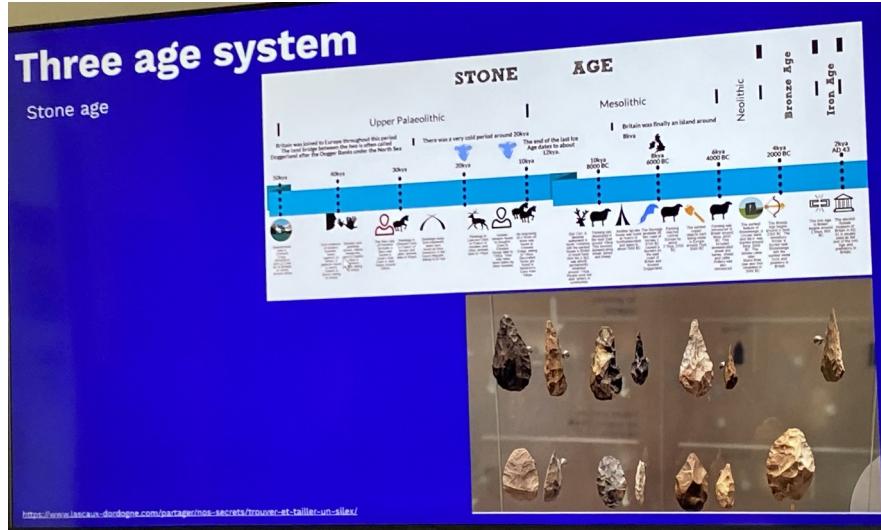
=> Étude des caractéristiques de la structure : La longueur de liaison (plus courte donc plus intense) explique la structure cristalline.

Les distorsions confirment leur intérêt spécifique.



2 – CONTEXTE HISTORIQUE

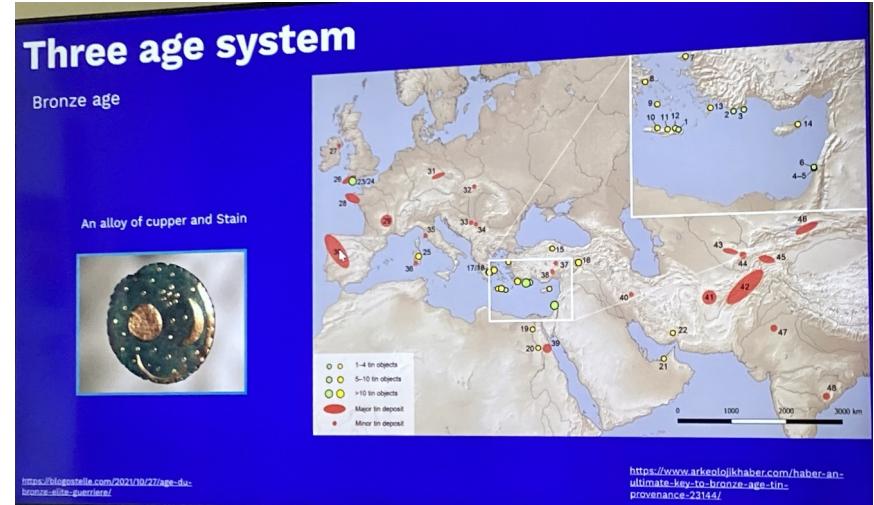
A chaque âge correspond un matériau.



L'âge se termine avec la découverte d'un autre matériau, développant de nouvelles connaissances.

L'âge de bronze favorise le commerce.

L'empire romain domine grâce à l'étain.

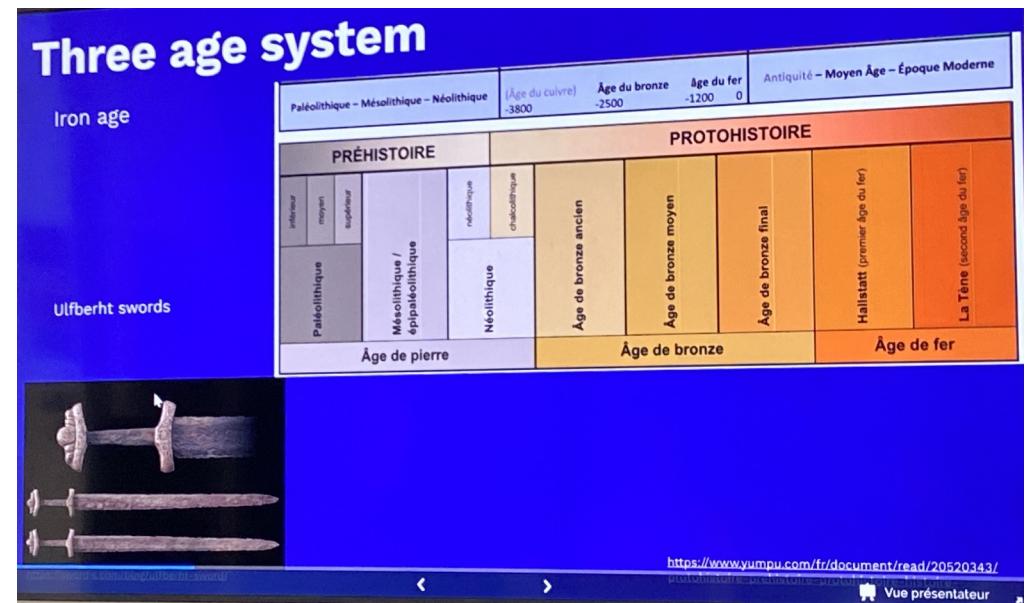
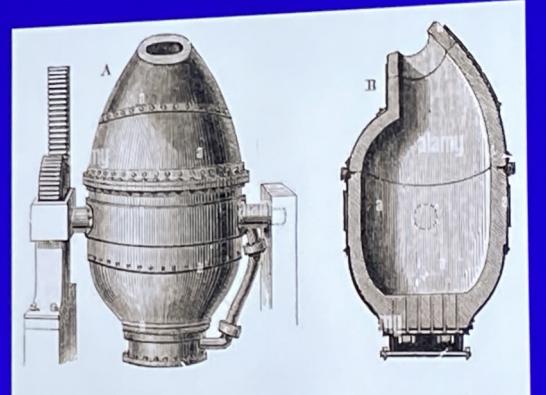


L'âge de fer est gage d'une position dans la société (secret du fabrication de l'épée se transmet de père en fils).

Material Discovery

Age of Steel

Bessemer process 1850



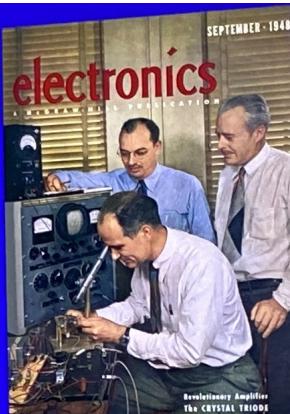
A l'âge du fer, le développement de la période industrielle et l'expansion des réseaux (chemin de fer...) font la richesse de l'Angleterre.

Après l'âge du transistor...

Le semi-conducteur, avec la découverte du triode pour fabrication du transistor.

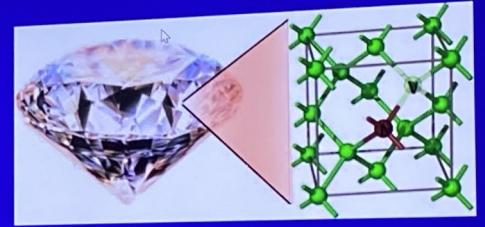
Material Discovery

Age of semi-conductors



Material Discovery

Age of diamant ?



From the Entropocene to Omnimatereocene ?

<https://e3.physik.tu-dortmund.de/cms/en/Team-Suter/Research-Interests/Diamond-NV-center/index.html>

...L'âge du diamant ?

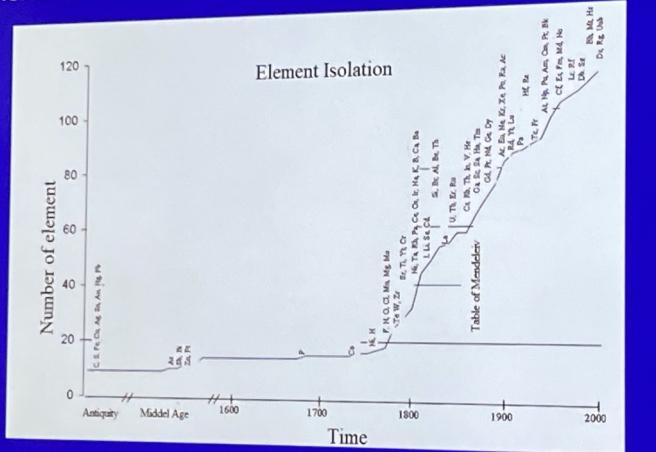
Différence au niveau des inclusions qui font la rareté du diamant naturel vs artificiel.

L'isolement d'un composant :

=> intérêt démarre avec Mendeleïv (tableau qui reste à compléter) et Lavoisier

Material Discovery

Isolation of elements from the Antiquity



https://fr.wikipedia.org/wiki/Histoire_de_la_d%C3%A9couverte_des_%C3%AEments_chimiques

Inventory of inorganic material landscape

Table 1. Estimates for the Number of Possible Inorganic Materials Allowing for Variable Oxidation States and Stoichiometry with the Constraints of Charge Neutrality and Electronegativity Balance

| Type | Constraint ^a | Number |
|-------------|-------------------------|-------------------|
| A_mB_n | - | 3,483,129 |
| A_wB_x | q | 58,614 |
| A_wB_x | $q + \chi$ | 14,721 |
| $A_wB_zC_y$ | - | 4,753,229,039 |
| $A_wB_zC_y$ | q | 174,081,685 |
| $A_wB_zC_y$ | $q + \chi$ | 32,157,899 |
| $A_wB_zC_z$ | - | 4,139,315,402,300 |
| $A_wB_zC_z$ | q | 267,381,955,246 |
| $A_wB_zC_z$ | $q + \chi$ | 32,381,953,858 |

^a q, charge neutrality; X, electronegativity balance.

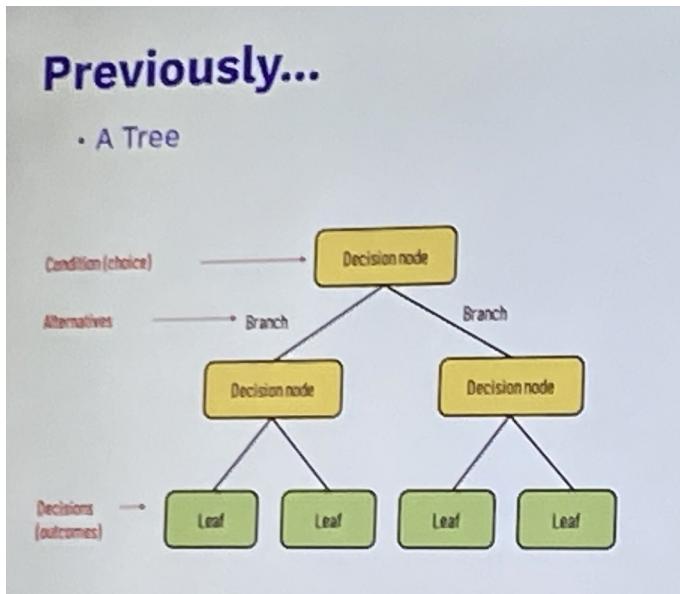
Davies, D. W. et al. Computational screening of all stoichiometric inorganic materials. Chem

La charge augmente et en fonction du type suivant règles de neutralité et négativité de charge

3 – GAME CHANGER

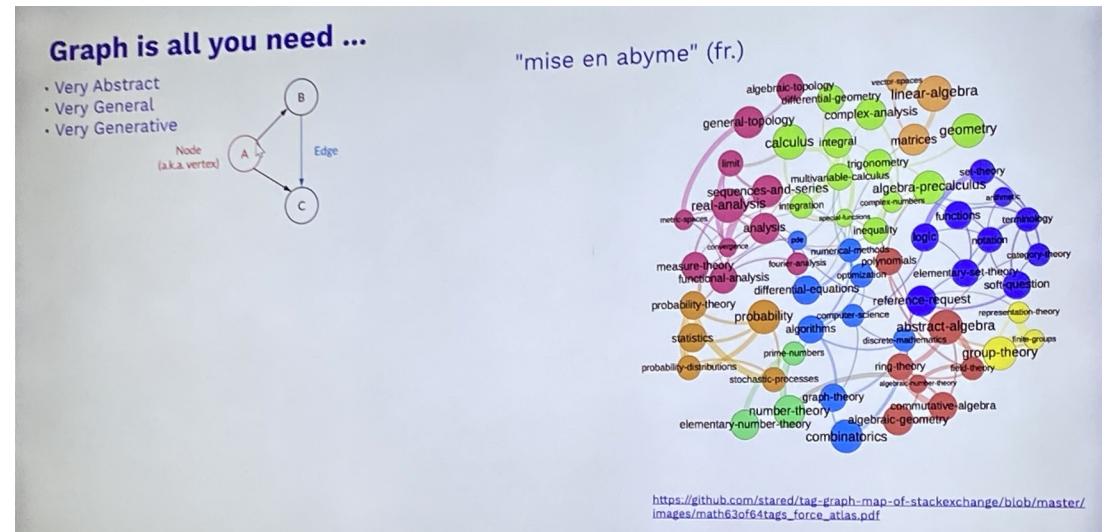
L'attention est tout ce qu'i vous faut.

L'abstraction est la force du graph et fait la force générative du Vertex



Arbre décisionnel (iris) – les nœuds

Relation entre point A et B.

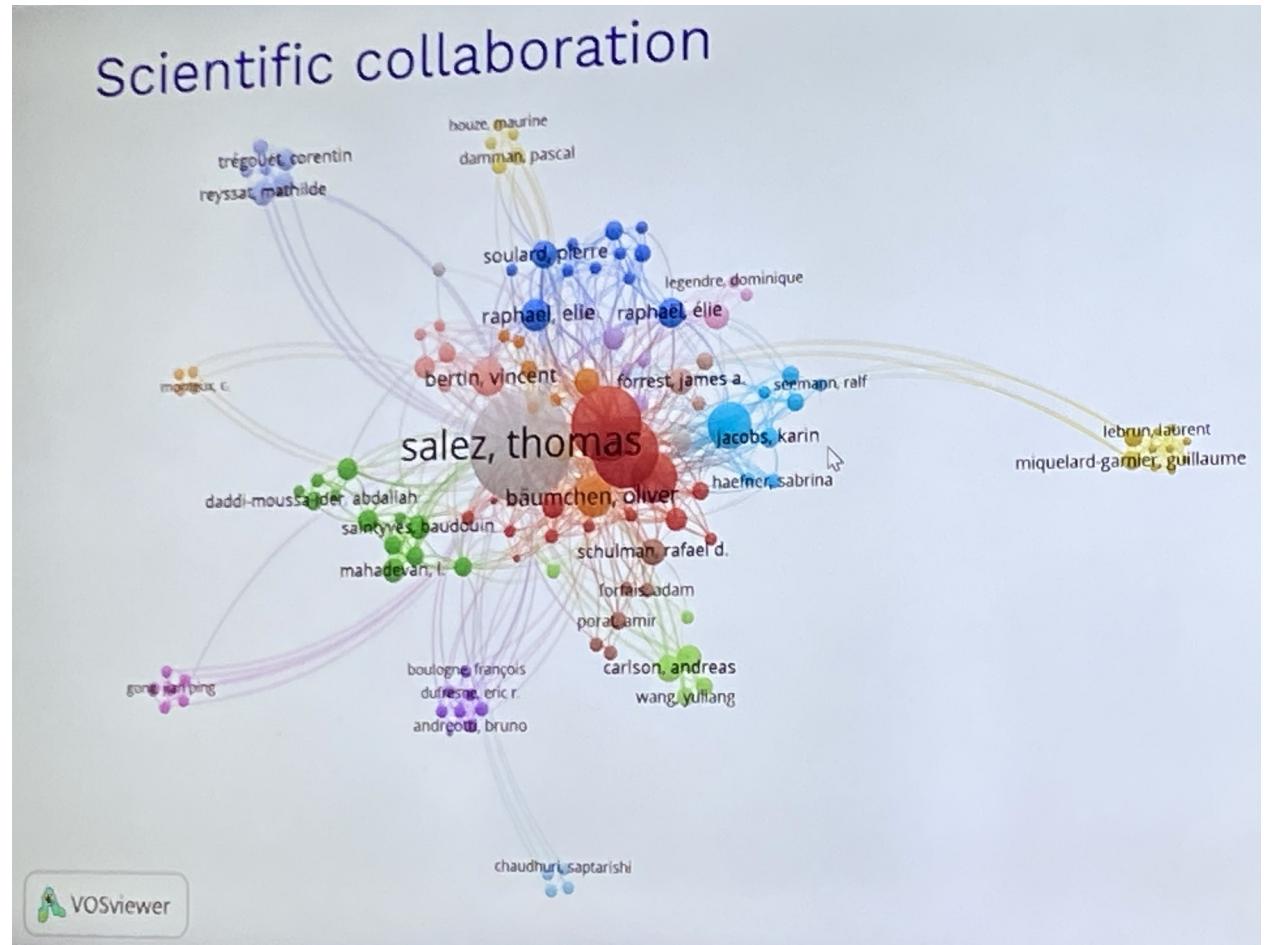


Théorie de graphe qui entretient des relations avec des combinatoires qui expliquent la génération de nouveaux cristaux, théorie des anneaux et topologie via lien de niveau qui explique les paysages.

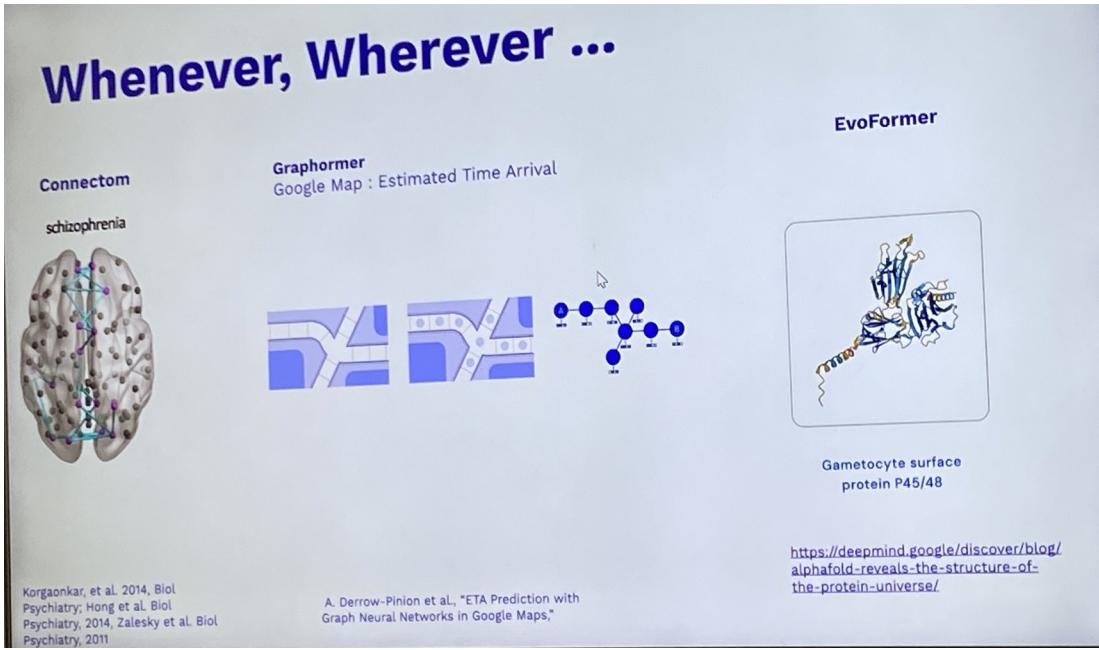
Scientific collaboration

Intensité explique les clusters qui façonnent la topologie

(outil : VOSviewer, en accès libre, pour visualiser l'avancement du travail entre les groupes de chercheurs)



Whenever, Wherever ...



Utilisation dans la recherche médicale pour trouver le pattern du graph expliquant la maladie (schizophrénie)

=> identifier la région du cerveau la plus sollicitée.

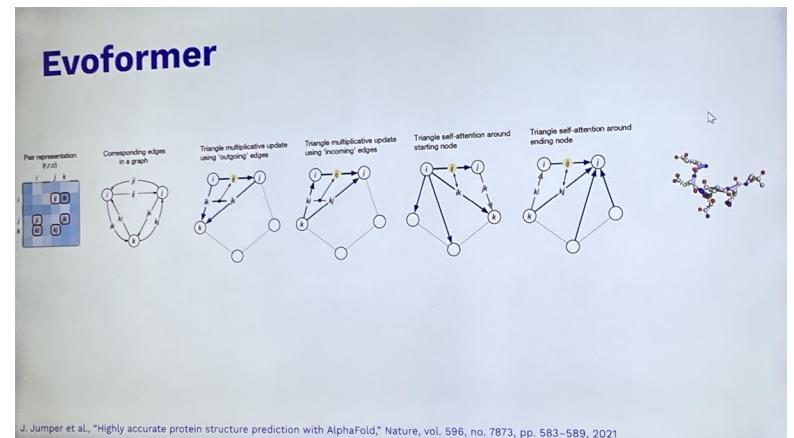
2018 – Google :

Comment passer de la séquence d'acide aminé à la structure de la protéine (evoform) ?

Les chercheurs ont « tué le game » pour s'attaquer à des complexes de protéines.

=> Variabilité entre anticorps et antigènes reste la difficulté à résoudre (configuration spatiale).

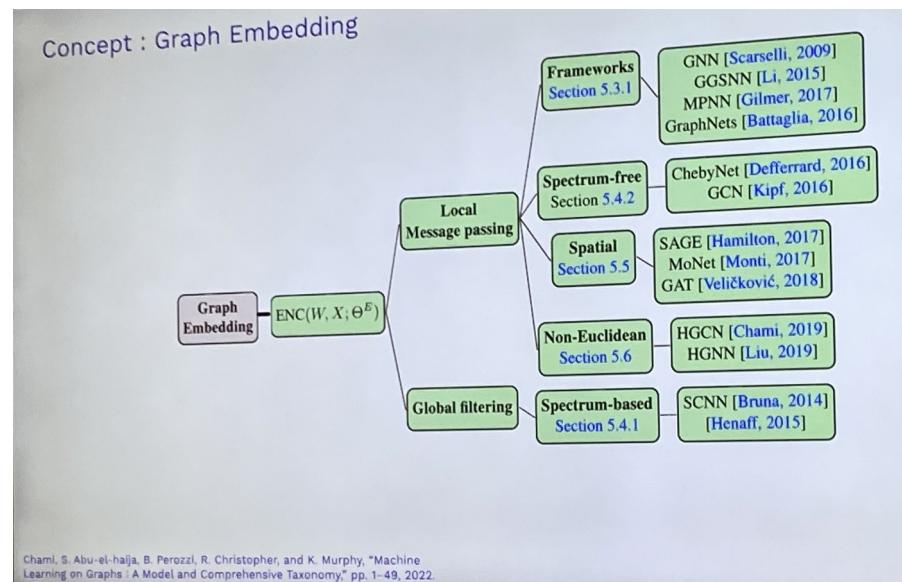
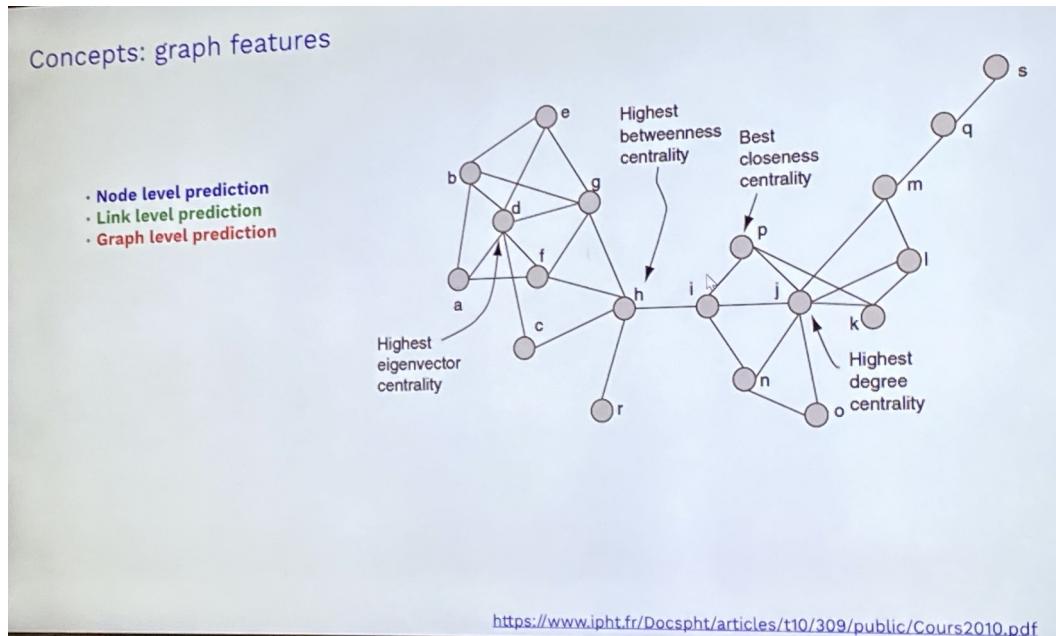
Prouesse Alphaform 2 et 3



Passer du game à la santé (structure de la protéine)

En cybersécurité :

Prédiction au niveau des nœuds, des liaisons (Edge) des clusters, permet d'éviter les attaques informatiques afin d'éviter que le réseau ne soit fragile.



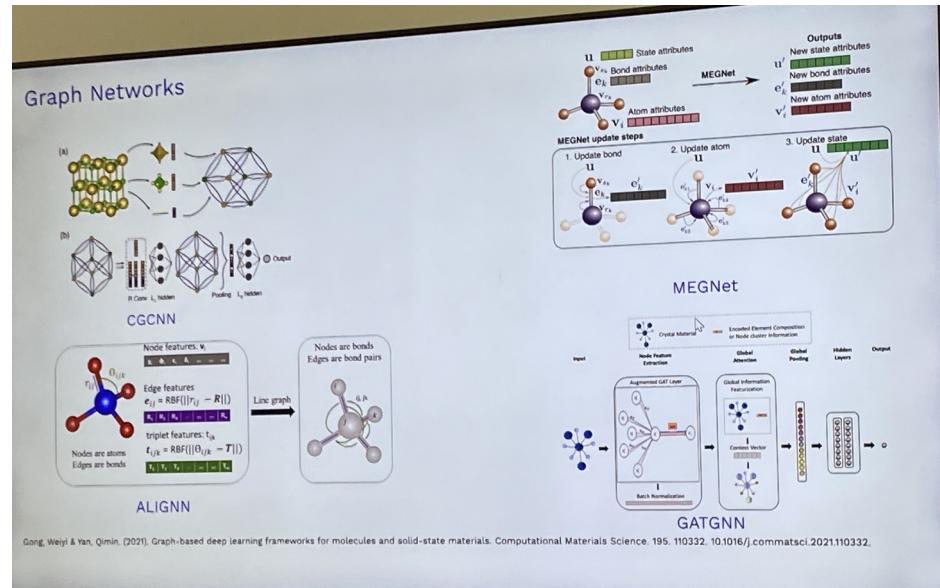
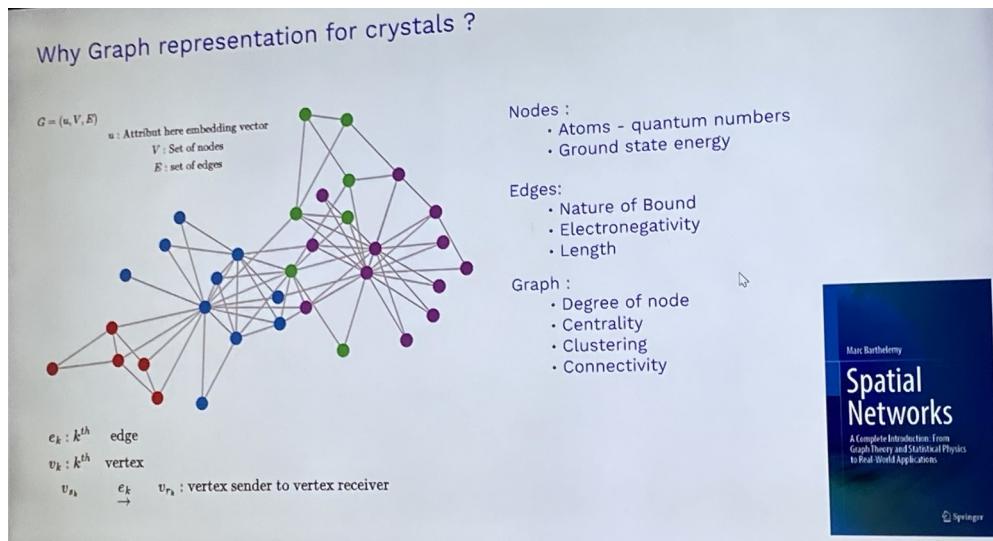
Article taxonomie :

Encapsulation du graph permet l'apparition du passage de message.

2013/2014 représente l'état de l'art dans le domaine de représentation de graph. Approche non-euclidienne qui trouve son intérêt dans les graph.

Usage industriel reste limité (pharma, jeu de foot, cybersécurité, finance avec éclatement d'une bulle ou apparition du cygne noir...) depuis 2017.

Possibilité de faire des apprentissages sur le degré de chaque nœud par l'étude des cristaux.



Utilisation propre pour chaque encapsulation de la connaissance d'un atome par rapport à son voisinage immédiat.

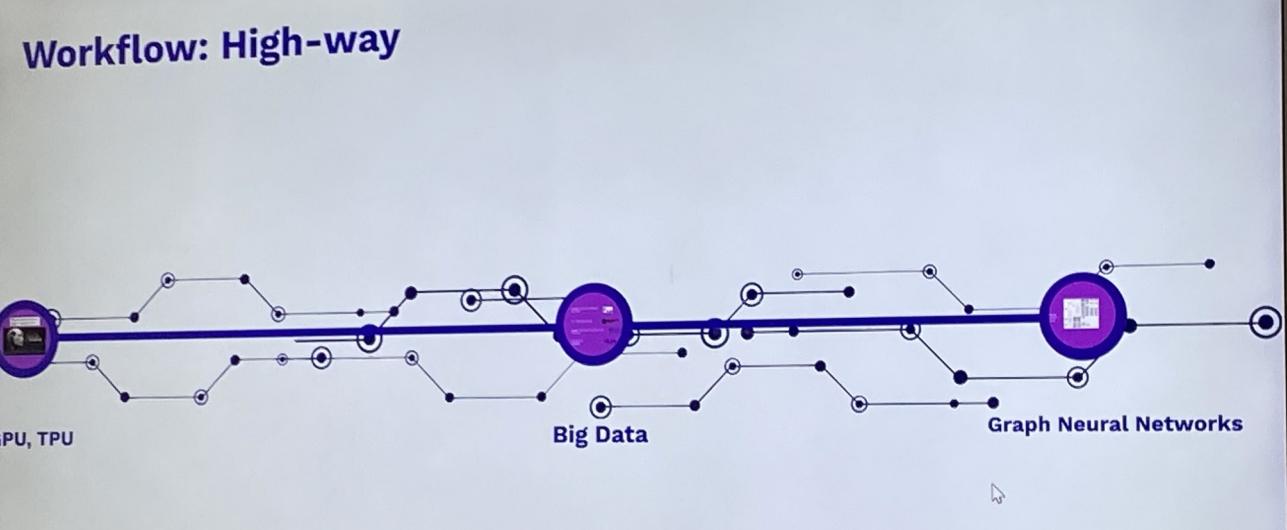
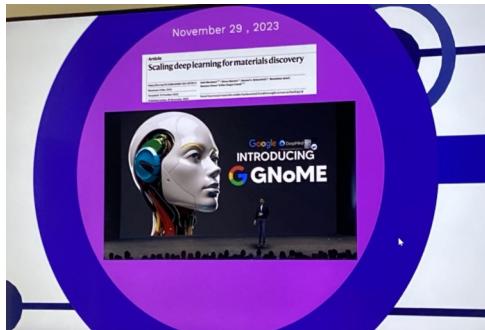
Alignement des planètes GPU et DPU

(gestion des tâches graphiques complexes et centres de données).

Optimisation :

- de la gestion des informations
- du renforcement de la sécurité,
- de l'efficacité du stockage
- et de la transmission des données.

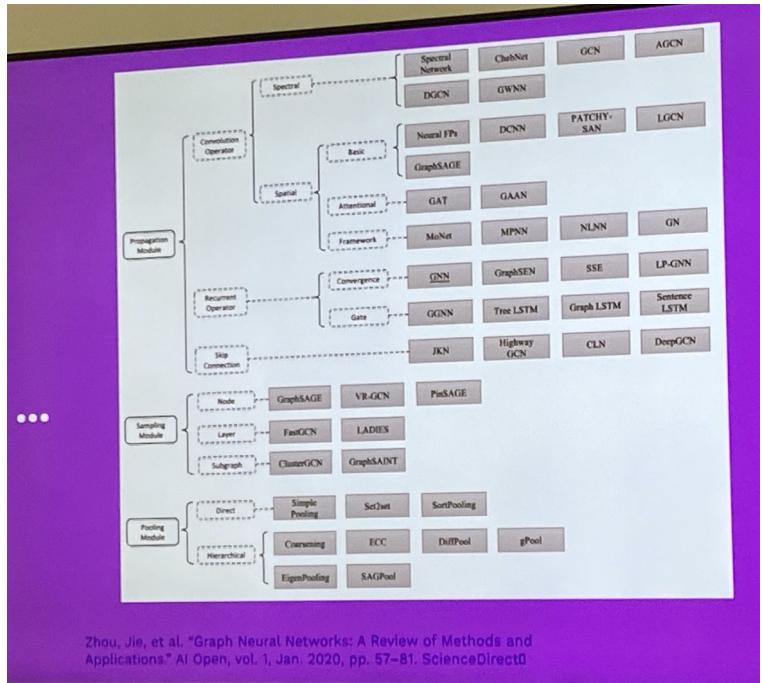
Avec la cheville ouvrière de DeepMind...



... Nouvelle façon de faire du produit matriciel.

Data Base :

Données accessibles avec un simple profil pour entraîner et aboutir au GNoME.



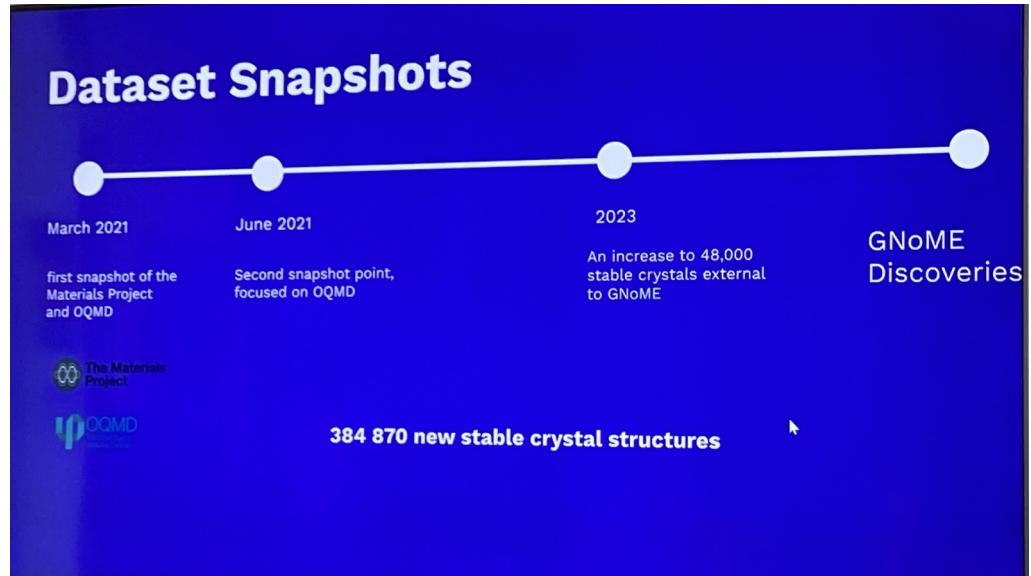
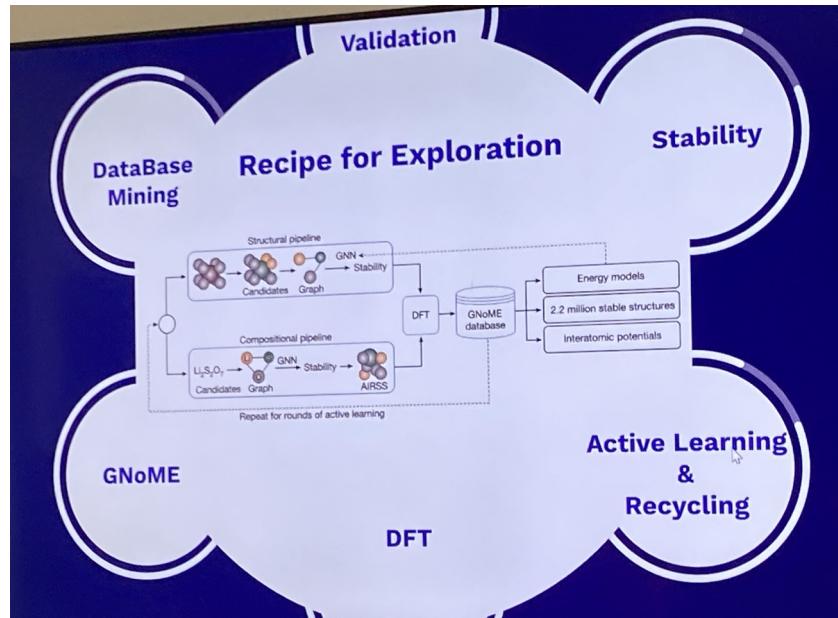
Zhou, Jie, et al. "Graph Neural Networks: A Review of Methods and Applications." *AI Open*, vol. 1, Jan. 2020, pp. 57–81. ScienceDirect®

Classement des GNM :

Alignement des graphes permet de faire des calculs parallèles.

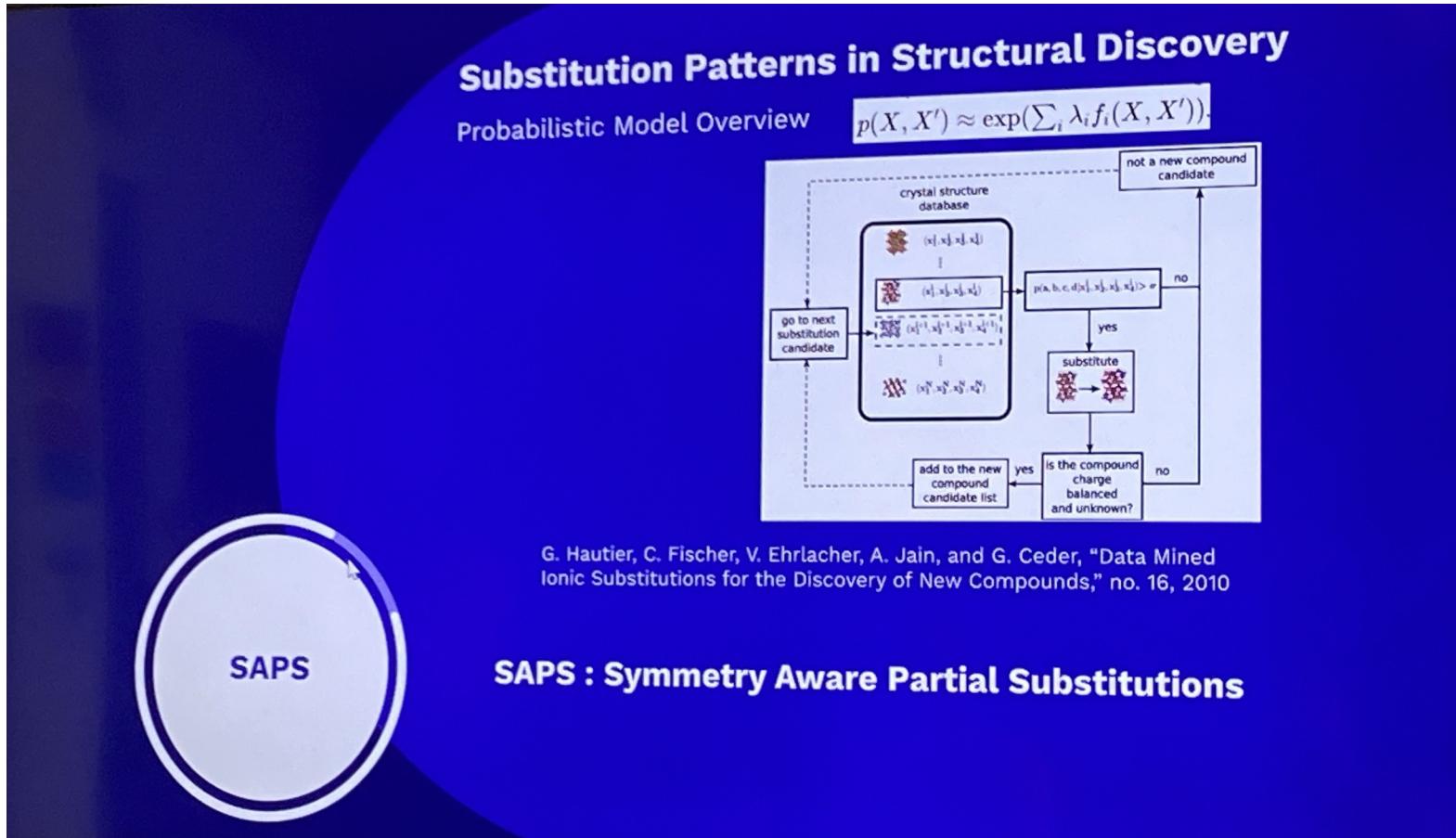
4 – SOUS LE CAPOT

Recherche orientée



Base de données monumentale (384 870 nouvelles structures du binaire à l'hexanaire) du fait des différentes propriétés (rapport de 1 million entre ce type de donnée et transcription textes).

Travail sur schéma de probabilités conditionnelles amorcé par Geoffrey HAUTIER en 2010 (thèse de doctorat).



Partir d'un composant X qui répertorie les composants.

La probabilité a plusieurs variables.

1. Probabilistic model for ionic substitution

- Likelihood of ionic substitutions using a probabilistic framework based on experimental crystal structure
- assesses the likelihood that two compounds can exist with the same crystal structure after substitution of their ions

Compound : n-Vector

$$X = (X_1, X_2, \dots, X_n)$$

Assesses the likelihood of these compounds coexisting in the same crystal structure.

$$p_n(X, X') = p_n(X_1, X_2, \dots, X_n, X'_1, X'_2, \dots, X'_n)$$

Multivariate probability function is approximated using feature functions

Entraînement de l'approximation sur la fonction de ces points

2. Feature Functions

f are binary indicators that determine whether two ions can substitute for each other under specific conditions

$$f(X, X') = \begin{cases} 1 & \text{if } \text{Ca}^{2+} \text{ substitutes for } \text{Ba}^{2+} \\ 0 & \text{otherwise} \end{cases}$$

Approximation : By summing the weighted feature function f

$$p_n(X, X') \approx e^{\sum_i \lambda_i f_i(X, X')}$$

λ_i are the weights associated with the feature functions f_i .
 Z is a partition function that ensures the normalization of the probability

3. Binary Feature Model

The model assumes that substitution rules are binary only pairs of ions

$$f_{a,b}(X, X') = \begin{cases} 1 & \text{if } X_k = a \text{ and } X'_k = b \\ 0 & \text{otherwise} \end{cases}$$

The model does not account for the entire chemical context

4. Training the Model

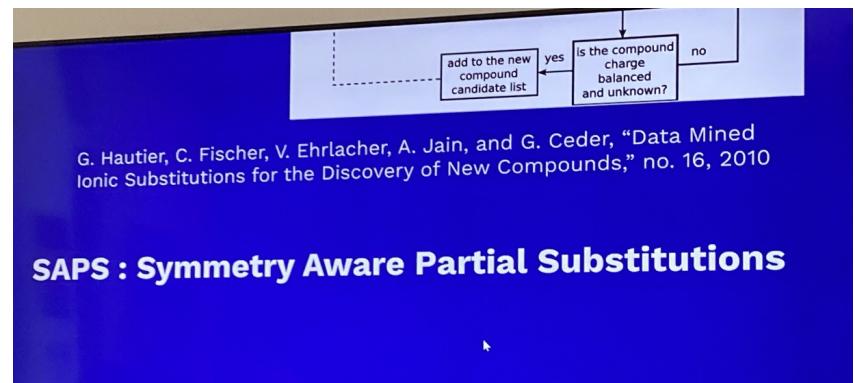
The weights are trained using database (ICSD)

Maximizing the log-likelihood of observing the training data

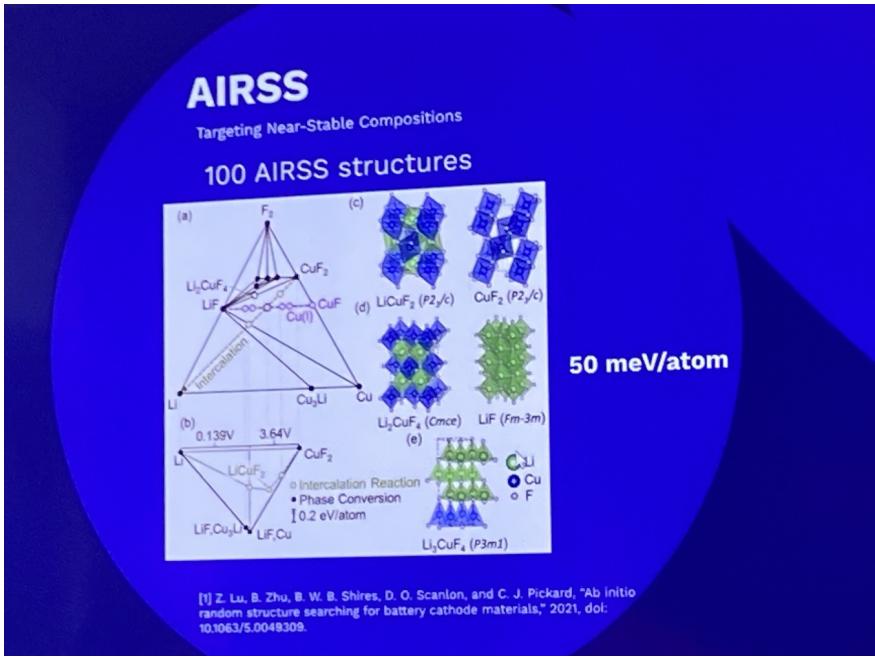
$$l(D, \lambda) = \sum_{t=1}^m \log p((x, x')_t | \lambda)$$

find the set of weights that maximize this log-likelihood

Fonction indicatrice.
 On peut changer le Calcium par du Barium.



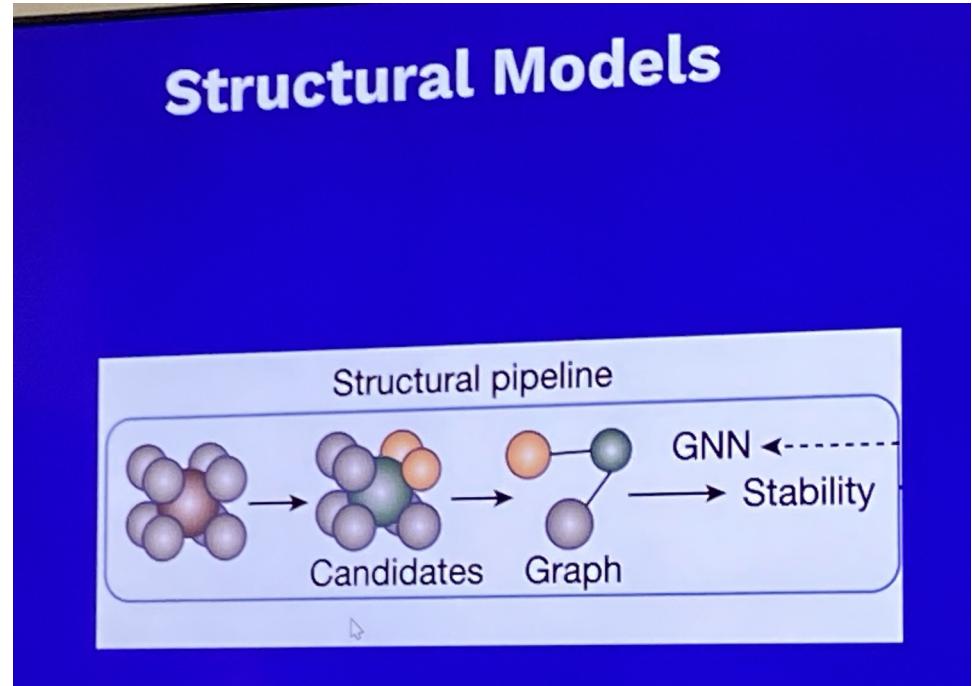
Nota : Il existe des cas où ça ne marche pas (structure non modifiable).



Énergie de formation de 50 meV/atom

Critique de sélection en fonction de l'énergie de formation (ADFT), calculs puissants faits dans la boucle.

=> Règle de base : Un variant par permutation.



Approche double, structurelle et composition.

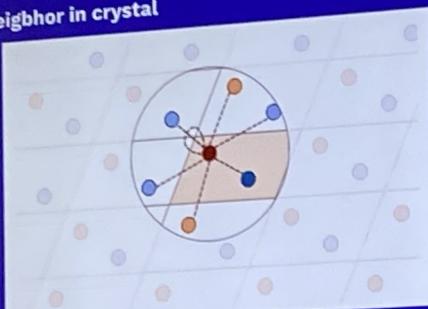
Apprentissage par représentation, avec vecteurs de base, ensuite fait sur liaisons par encapsulation.

Il n'y a que le passage de message qui va jouer. Apprendre sur le voisin pour connaître l'état de liaison, puis agrégation, puis un autre vecteur par sommation. Enfin, encapsulation de l'ensemble.

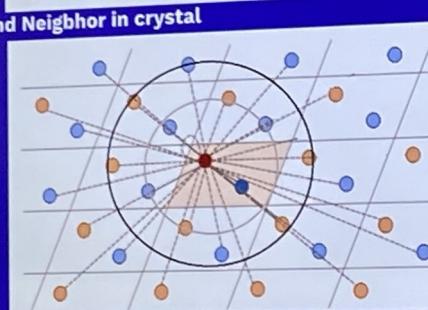
Dilution dans la propagation au fur et à mesure de l'élargissement des cercles construits par connexion.

How it works

• First Neighbor in crystal



• Second Neighbor in crystal



W. L. Hamilton, "Graph Representation Learning," vol. 14, no. 3, 2020.

Basic GNN message passing

$$h_u^k = \sigma \left(W_{self}^k h_u^{(k-1)} + W_{neighbor}^{(k)} \sum_{v \in N(u)} h_v^{(k-1)} + b^{(k)} \right)$$

Where:

W_{self}^k : Self trainable parameter matrices

$W_{neighbor}^k$: Neighbor trainable parameter matrices

σ : elementwise non-linearity : ReLU, tanh

La non-linéarité a la clé :

Si application aux communautés pour contrôle des individus/populations
=> risque exponentiel / mesures liberticides.

Architecture d'apprentissage de message.

Apprentissage crucial suivant fonction (algorithme) très simple par permutation, selon poids des nœuds (explique la stabilité d'une structure comparée à une autre), agrégation puis perception.

Inside MPL : Architecture

$$h_u^{(t+1)} = (\text{UPDATE})^{(k)}(h_u^{(t)}, \text{AGGREGATE}^{(k)}, (h_v^{(t)}, \forall v \in N(v)))$$

1. Node Feature Update

$$h_v^{t+1} = U_t(h_v^t, m_v^{t+1})$$

$$h_v^{t+1} = \text{ReLU}(W_1 h_v^t + b_1)$$

2. Aggregation function

$$\text{Summation } m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{vw})$$

$$\text{Mean Pooling } m_v^{t+1} = \frac{1}{|N(v)|} \sum_{w \in N(v)} h_w^t$$

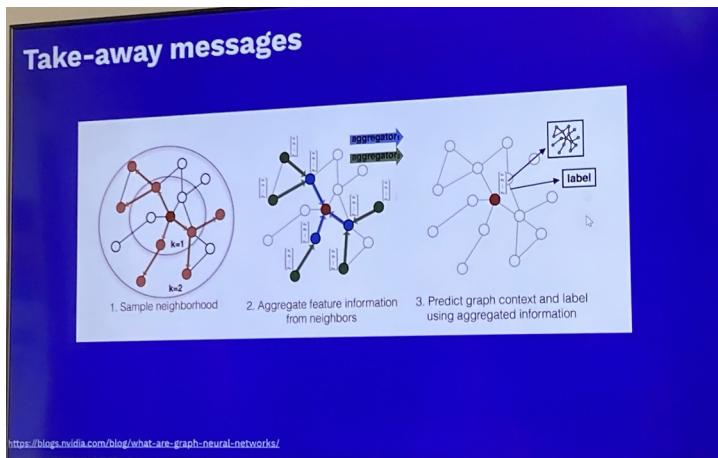
3. Readout = Global Pooling: Multi-Layers Perceptron (MLP)

The global graph representation is passed through a Multi-Layer Perceptron (MLP) to predict a scalar property : the formation energy of the material.

$$m_{N(u)} = \text{MLP}_\theta \left(\sum_{v \in N(u)} \text{MLP}_\phi(h_v) \right)$$

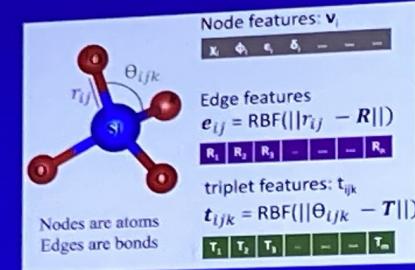
Apprendre de la représentation en cours d'étude avant de poursuivre au second cercle.

Après plusieurs essais, calcul sur la représentation finale.



Training process : working exemple SiO4

• Graph-structured data



• Graph Featurization

• Message Passing Learning : 4 layers updated through layers. This is done using various update functions

• Residual Connections : address the vanishing gradient

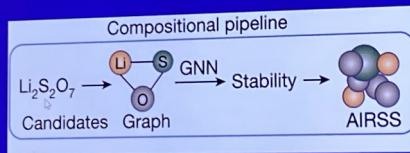
Passage de messages utilisé 4 fois.

Effet délétère du radian est atténué par l'emploi de l'adresse résiduelle de connexion. C'est là que les chercheurs font le **salto arrière** pour comprendre et enfin publier.

(BM sur vente des GPU).

Compositional Models

- creating a custom training set by running **AIRSS** simulations on novel compositions
- Focuses on materials' chemical compositions to predict their physical, chemical, or structural properties.



Contraintes du cristal : manque de transparence

Inspiration d'un modèle de représentation graph de stoechiométrie (bonne proportion) Roost.

Difficulté réside dans la stokiostoechiométrie ; C'est la base de la recherche + graphes d'attention.

The screenshot shows a research article from *nature communications*. The title is "Material symmetry recognition and property prediction accomplished by crystal capsule representation". The authors are Chao Liang¹, Yilin Li², Roushong², Caiyan Ye², Chong Li¹, Bao Wang¹, and Hua Chen¹. The article was received on 10 June 2023 and accepted on 7 August 2023. The abstract discusses a model that captures equivariant transformations that preserve spatial relationships between atoms. Below the abstract, there are two diagrams illustrating crystal symmetry. The left diagram shows a central atom A3 surrounded by other atoms, with dashed lines indicating symmetry operations like Translation, Rotation, and Reflection. The right diagram shows a more complex crystal structure with similar symmetry analysis.

Fait en août 2023 – travaux sur la représentation des symétries (awarness).

Tour de force :
deviner la symétrie (torsion, miroir...) + calcul quantique qui demande beaucoup de puissance énergétique.

Formation energy

$$\Delta H_f = E_{\text{total}}(A_xB_y) - (x \cdot E_A + y \cdot E_B)$$

Where:

- $E_{\text{total}}(A_xB_y)$ is the total energy of the compound A_xB_y (calculated, for example, using Density Functional Theory (DFT)).
- E_A and E_B are the energies of the elements A and B in their reference states (usually in their pure form in the most stable state).
- x and y are the stoichiometric coefficients in the compound A_xB_y .

Il faut un résultat négatif (soustraction de l'énergie) pour donner des structures stables.

Active learning

The image consists of two parts. The top part is a presentation slide titled "Learning rate schedules". It contains the following text:

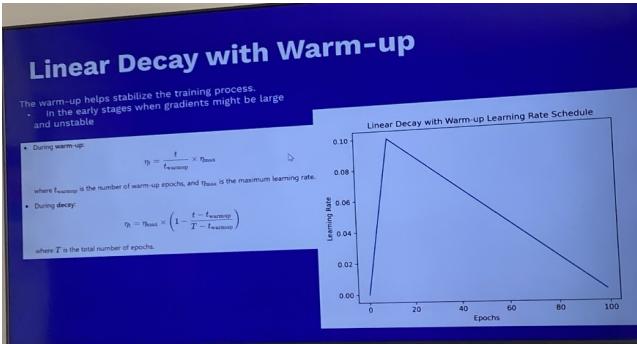
- Dynamically adjust the learning rate during training

A diagram shows a loop with a circle containing a dot and arrows forming a cycle. Below the diagram is the text "Repeat for rounds of active learning". To the right of the slide, there are two circular icons: one labeled "Cosine Decay" and another labeled "Warmup". The bottom part is a photograph of a man with glasses and a beard, wearing a dark blue shirt, standing in front of a whiteboard and gesturing with his hands. A laptop is on a table in front of him. The whiteboard displays the same "Learning rate schedules" slide from above.

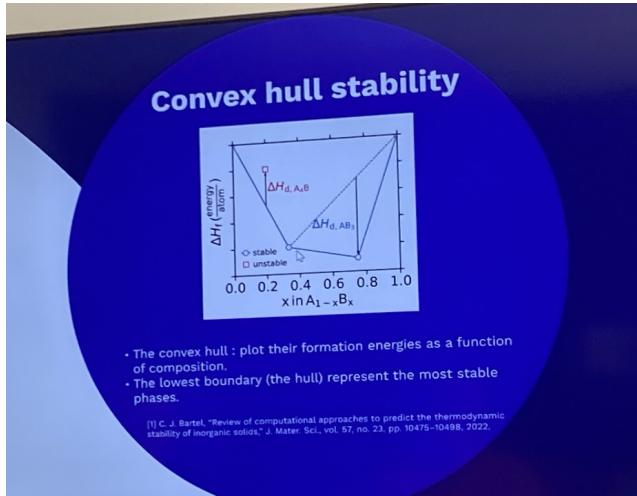
DLS - Meet-up du 17/09/2024

Réduction de l'erreur est grande avant le ré-entraînement (réguler le taux d'apprentissage par cosinus déc腺ant) ; n閙閞ise d'identifier l'erreur avant de quantifier le taux d'erreur (?)

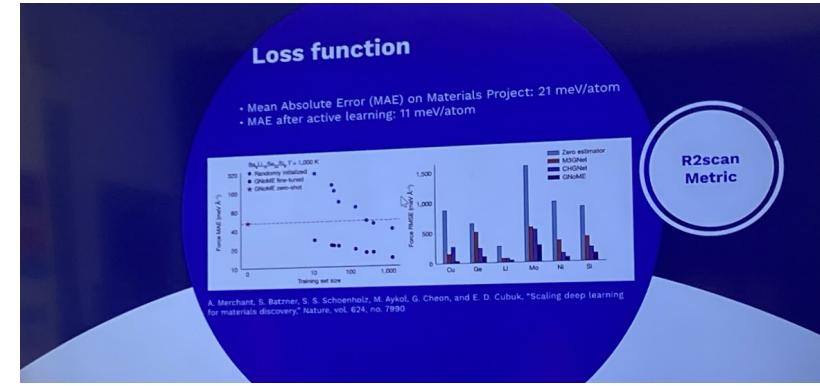
N閙閞ise d'avoir le potentiel pour chaque 脎lement de l'atome, donc avoir les moyens pour calculer.



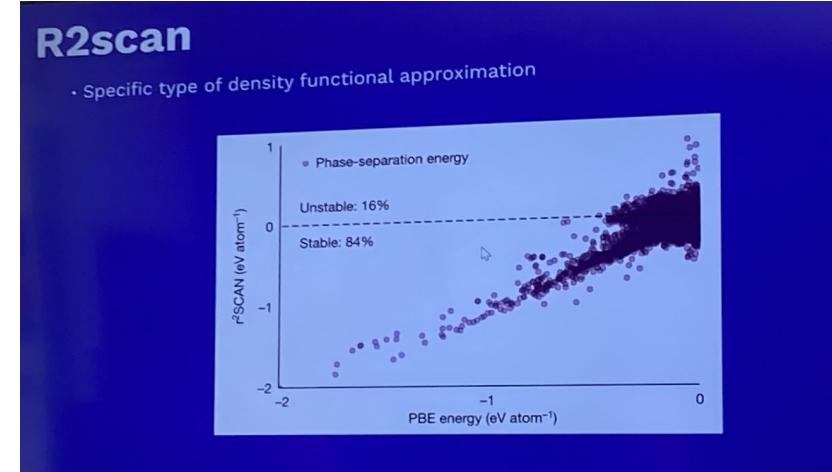
Régule le taux d'apprentissage suivant les époques et d'aller vite en direction de la bonne zone de l'espace.



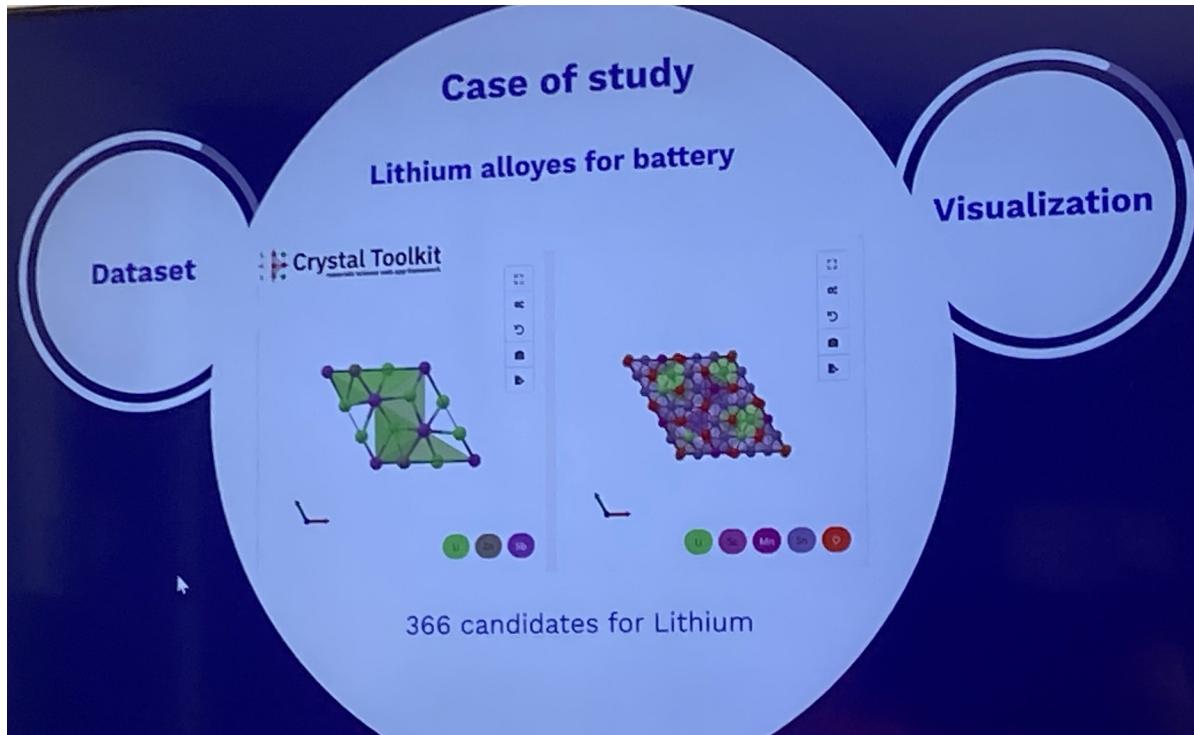
Enveloppe convexe :
contrôle de la stabilité de la structure (stable ou non).



Apprentissage actif (GnoME dans le violet).
Autre garde fou de la DFT est le R2scan pour calculer l'inertie de la structure



5 – ÉTUDE DE CAS



Lithium publication du Data set



Groupe ponctuel, groupe d'espace, système cristallin : éléments jamais vus, jamais synthétisés et le gap entre les bandes de valences et bandes électroniques.

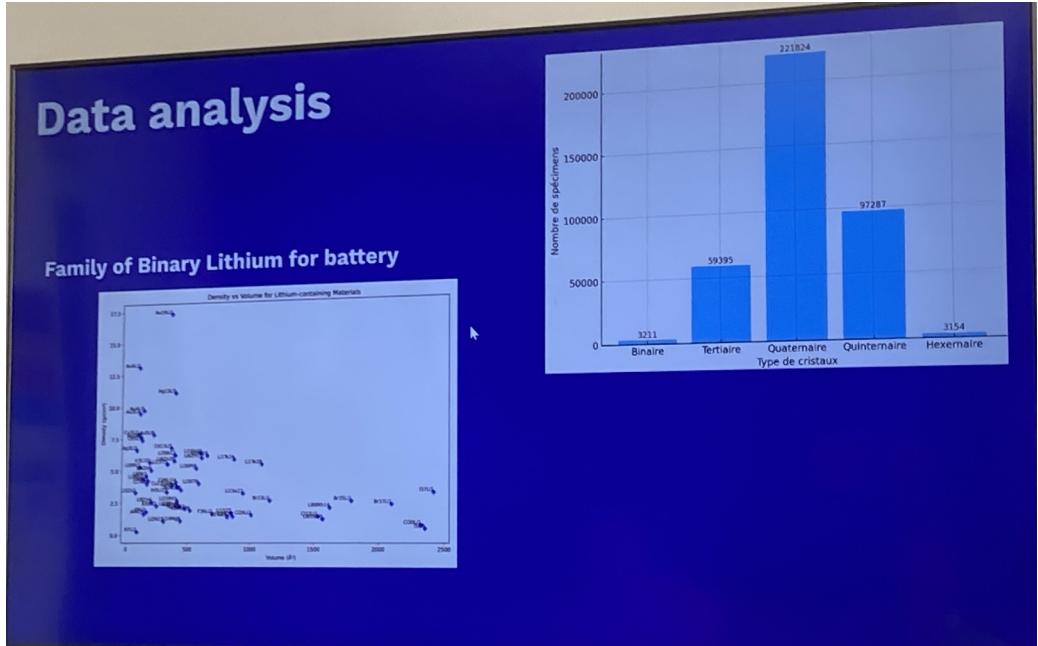
Structure : voir les binaires avec uniquement du Lithium.

Là de suite permet d'avoir des charges d'intensité plus rapides.

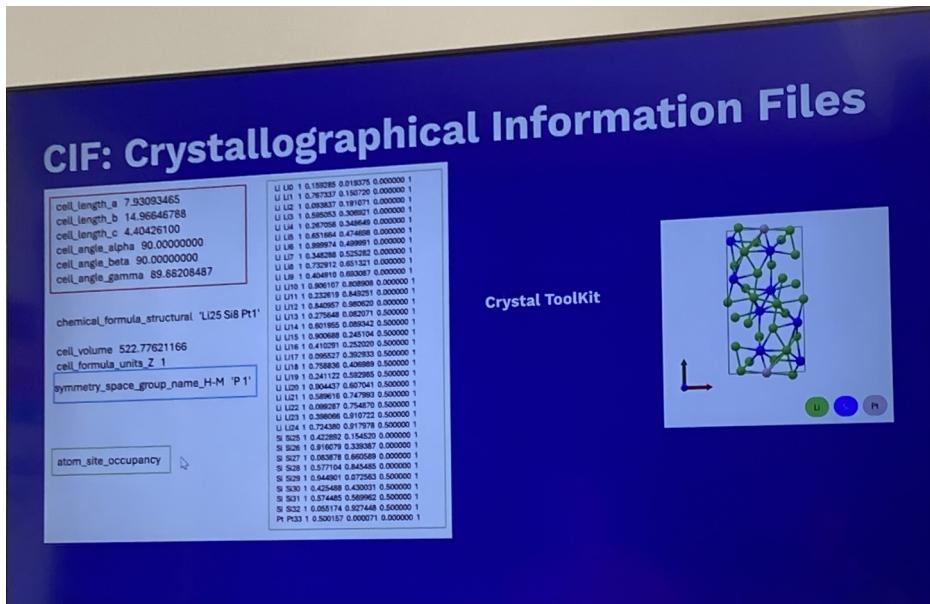
Structure variable suivant densité et volume qui conditionnent la puissance et efficacité des batteries (combinaisons improbables : Lithium + Mercure + Or ; Lithium + Hydrogène)...

Dans quelles conditions on peut avoir ça ? Prédiction reste difficile même si 380 000 ans d'avance gagnés selon les chercheurs.

Champs des possibles non exploré. 336 candidats pour le Lithium (application dans isolation bâtiment, textile, etc.).



CIF : Carte d'identité du cristal



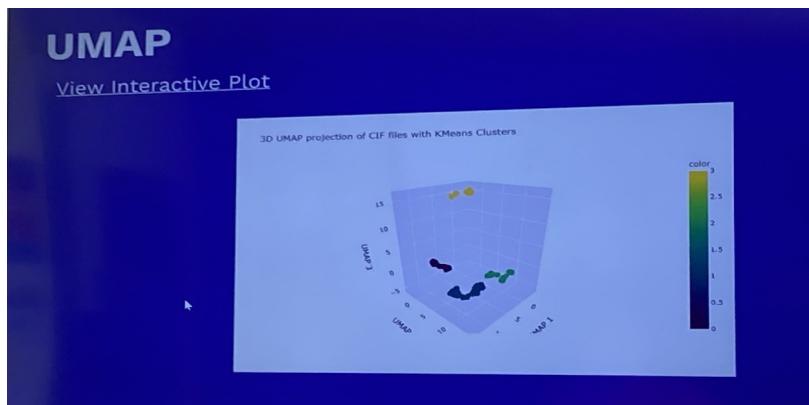
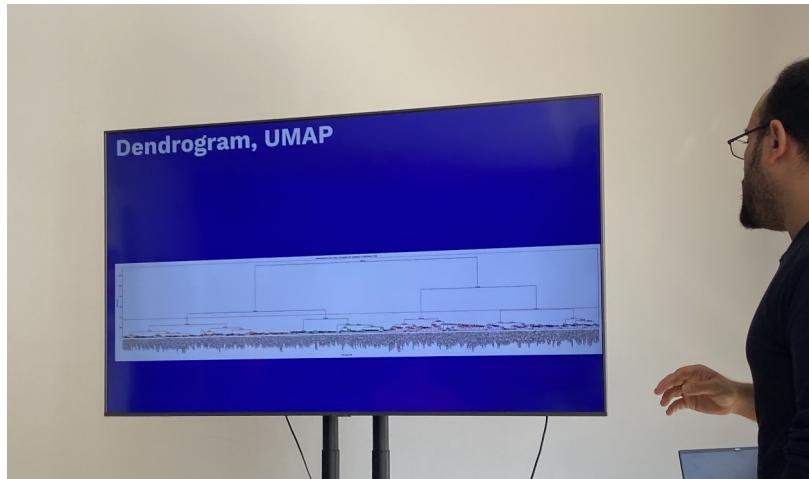
Mettre les atomes dans la bonne place pour que ça fonctionne. Travail incomplet.

Clusters existants ?

Suivant de variants de virus.

Connaissance réside au niveau du nœud.

Si embedding, séparation en 4 clusters. Éléments pittoresques se situent en périphérie.



6 – VALIDATION : Auto-LAb

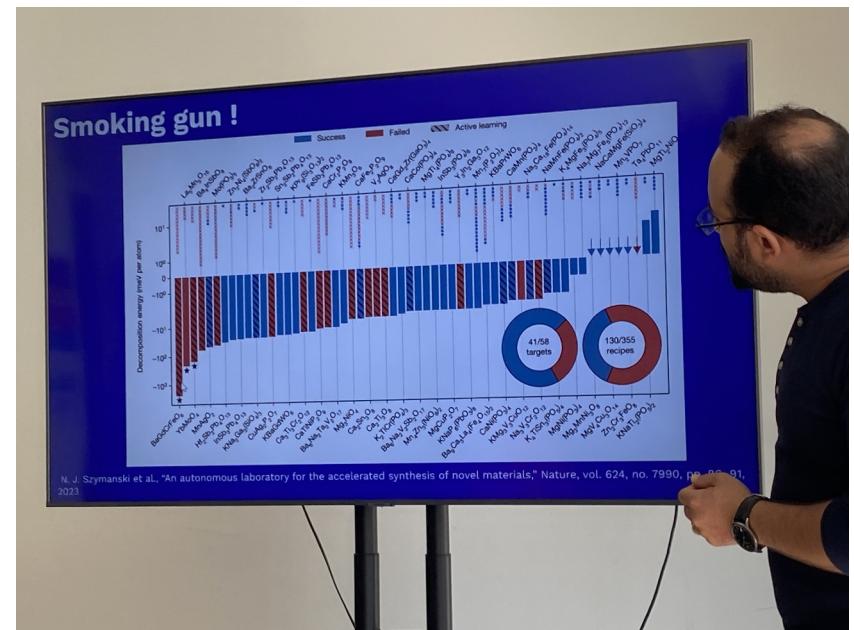
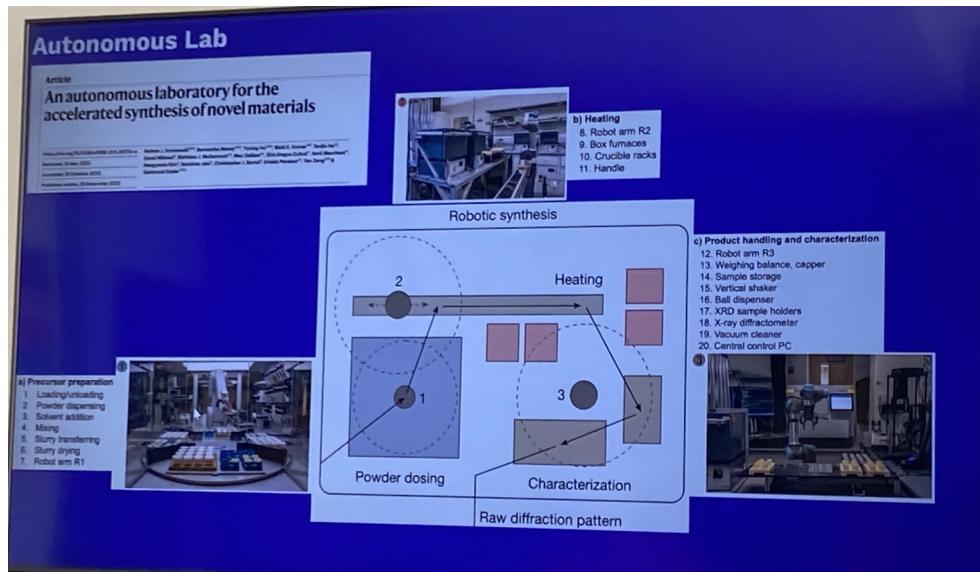
Autolab – SARCLAY

Activlearning pour deviner le chemin, l'étape par lequel le matériau va passer :

Phase des matériaux est différente suivant température, formation... Panel de 58 tests réussis (en bleu).

Deviner la recette (130 réussites sur 355), nécessite d'être capable de visualiser les graphes (trouver le chemin le plus court).

Sortir de la boîte noire.



Bibliography

Books

- Barthelemy, M. (2022). **Spatial Networks: A Complete Introduction: From Graph Theory and Statistical Physics to Real-World Applications.** Suisse: Springer International Publishing.

Websites

- Stanford CS224W: Machine Learning with Graphs | 2021 | Lecture 1.1 - Why Graphs
- <https://theaisummer.com/gnn-architectures/>
- <https://www.youtube.com/@TaylorSparks>

Articles:

- <https://www.ipht.fr/Docspht/articles/t10/309/public/Cours2010.pdf>
- **Graph Representation Learning Hamilton McGille, 2020**
- J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “**Neural Message Passing for Quantum Chemistry**.”
- I. Chami, S. Abu-el-haija, B. Perozzi, R. Christopher, and K. Murphy, “**Machine Learning on Graphs : A Model and Comprehensive Taxonomy**,”