altrove

Internship offer

contact : alessandro.cecchini@altrove.ai

October 17, 2024

Starting date : April-May
Duration : 6 months
Wage : 1600 euros
Location : Gentilly (metro line 14) 3 days per week / remote 2 days per week.

About altrove

- At Altrove we automate the synthesis process of novel inorganic materials with artificial intelligence, making the process 100x faster.
- Materials change the world. They are the cornerstone of smartphones, electric vehicles, and LEDs.
- We are a young and hungry startup (that recently raised €3.7M) that wants to change how materials are developed by combining AI and other computational approaches with lab operation.
- We are looking for driven people who will build fast and not hesitate to try novel approaches.

Generic Context

- Development and implementation of diffusion models / flow maps and other generative models within Altrove's state-of-the-art generative pipelines.
- Contributing to the optimisation of transformer systems associated with these pipelines.
- Performing detailed hyperparameter searches and optimising model performance.
- You will be working directly with our CTO and lead AI scientist (former MVA student).
- Computer resources (such as NVidia H200s GPUs) will be allocated to the interns.
- Altrove is a small company. That means that you will have considerable freedom and responsibility and get to work directly with everyone at the company.

Please, find beneath the description of the two internships.

If you want to apply to one of the topics please send an email with your CV to the email address displayed above. No need to hand a cover letter.

Topic 1: Generative modelling of chemical reactions

The recent Alphafold 3 model [Abr+24], uses a graph neural network diffusion model to predict small molecule / protein spatial conformation. At Altrove, we use similar models within inorganic chemistry, which present additional constraints, particularly that the final structures are periodic. This project will include development and implementation of a diffusion model [Kar+22] / flow maps [BAV24] capable of predicting inorganic crystal structures conditioned on reaction constraints.

The trained model will then be incorporated within a larger pipeline within altrove's tech stack. The goal is to be succesfully able to predict the likelihood of chemical reaction products given a set of starting materials (precursors) and metadata as conditional variables.

Prerequisites

- Knowledge of Jax (or PyTorch, but ideally Jax)
- Experience with Graph Neural Networks and Transformer Architectures
- Working knowledge of implementing diffusion models / flow maps (or other generative models)
- Background in Bayesian statistics.
- Understanding of multimodal AI architectures.
- Communication skills in English

Topic 2: Encoding crystal disorder within generative model pipelines

Altrove already has a diffusion model [Kar+22] pipeline that diffuses crystal structures given other constraints. However, such models are limited in diffusing to 'perfect' crystals. In reality, all crystal structures have some form of defects and disorder: some atoms might be missing 10-20% of the time, and atoms might be swapped with atoms of other types on occasion. These types of disorder can not be currently dealt with as the crystal structure representations for generative models are lacking. However, this disorder can be vital for material properties and stability.

The goal of this topic is to develop a representation that can both:

1. be plugged into a diffusion model pipeline
2. incorporate defects and disorder within the representation.

The work representations will be subsequently used in diffusion model architectures. Towards the end of the project, the plan is to generate novel disordered structures which will be tested in Altrove's lab to create a new material.

Prerequisites

- Knowledge of Jax or Pytorch
- Working knowledge of diffusion models / flow maps (or other generative models)
- Background and expertise in types of Fourier and other signal/frequency transforms.
- Math and signal processing background
- Communication skills in English

Candidate profile

We are looking for highly motivated students, finishing their Master 2 year specialized in artificial intelligence, statistics or mathematics. The ideal candidate would have already been exposed to most of the following topics :

- Graph Neural Networks (GNNs)
- Generative models such as diffusion models or flow maps
- Transformer architectures

Knowledge in Bayesian statistics, stochastic calculus, functional analysis and algebra is appreciated. Knowledge in chemistry is not required but is a plus.

The student will have experience in implementing artificial neural networks with either Pytorch or Jax (optimally Jax). Experience with Github is a plus.

References

[Kar+22] Tero Karras et al. "Elucidating the Design Space of Diffusion-Based Generative Models". In: *Proc. NeurIPS*. 2022.

[Abr+24] Josh Abramson et al. "Accurate structure prediction of biomolecular interactions with AlphaFold 3". In: *Nature* 630.8016 (June 2024), pp. 493–500. DOI: `10.1038/s41586-024-07487-w`. URL: `https://doi.org/10.1038/s41586-024-07487-w`.

[BAV24] Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. *Flow Map Matching*. 2024. arXiv: `2406.07507 [cs.LG]`. URL: `https://arxiv.org/abs/2406.07507`.

Appendix: Inorganic Crystal Description

A crystal is composed of an infinite number of atoms' types $\mathcal{A}$ lying in $\mathbb{R}^3$ at positions $\mathcal{X}$ bounded together by electromagnetic forces. The atoms' positions respect a certain number of symmetries, that is, to every crystal graph we can associate a *space group* $\mathcal{G}$ which acts on $\mathbb{R}^3$ and leaves $(\mathcal{A}, \mathcal{X})$ invariant. An element $g = (\mathbf{W}, \mathbf{v})$ of a space group $\mathcal{G}$ is decomposed into a translation $(\mathbb{1}_{\mathcal{G}}, \mathbf{v}) \in \mathcal{T}$, where $\mathcal{T}$ is the translation subgroup of $\mathcal{G}$ and an orthogonal operator $W \in \mathcal{P}$ where $\mathcal{P}$ is the *point group* which is the subgroup of the orthogonal group $O(3)$. The translation subgroup is associated with the translation lattice $\mathbb{L} = \{\mathbf{v} \in \mathcal{T} : (W, \mathbf{v}) \in \mathcal{G}\}$. From group theory we know that there exists three linearly independent lattice basis vectors $\mathbf{L} = (\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3) \in \mathrm{Mat}_{3,3}(\mathbb{R})$ where coordinates are given with respect to the canonical basis, which describe all translation leaving the crystal pattern invariant through integral linear combinations of those, that is

$$\mathbb{L} = \bigcup_{\mathbf{k} \in \mathbb{Z}^3} \mathbf{L}\mathbf{k}$$

Furthermore the point group $\mathcal{P}$ acts on the translation lattice, i.e. if $\mathbf{W} \in \mathbb{P}$ then

$$\forall \mathbf{v} \in \mathbb{R}, \mathbf{v} \in \mathbb{L} \implies \mathbf{W}\mathbf{v} \in \mathbb{L}$$

.

A *primitive unit-cell* is the parallelepiped spanned by the basis vector of the lattice, i.e.

$$\mathcal{C}_{\text{cell}} = \bigcup_{\mathbf{s} \in [0,1)^3} \mathbf{L}\mathbf{s}$$

The set of all translations of $\mathcal{C}_{\text{cell}}$ by $\mathbf{v} \in \mathbb{L}$ is a partition of $\mathbb{R}^3$. Hence, given N_{cell} atoms lying in the primitive cell, a crystal graph is compactly summarized as a triple $(\mathbf{A}_{\text{cell}}, \mathbf{X}_{\text{cell}}, \mathbf{L})$ where

- $n \in \{1, ..., N_{\text{cell}}\}$ is a chosen index
- $\mathbf{A}_{\text{cell}} = \{a_1^{\text{cell}}, ..., a_{N_{\text{cell}}}^{\text{cell}}\} = \mathcal{A} \cap \mathcal{C}_{\text{cell}}$ (by abuse of notation), $a_n^{\text{cell}} \in \{1, ..., 118\}$ represents the atom type with respect to periodic table numbering.
- $\mathbf{X}_{\text{cell}} = \{\mathbf{x}_1^{\text{cell}}, ..., \mathbf{x}_{N_{\text{cell}}}^{\text{cell}}\} = \mathcal{X} \cap \mathcal{C}_{\text{cell}}$ and $\mathbf{x}_n^{\text{cell}} \in \mathbb{R}^3$ represents the coordinate of the atom with respect to the canonical basis. The fractional coordinate $\mathbf{q}_n^{\text{cell}} \in [0,1)^3 \cap \mathbb{Q}^3$ with respect to lattice basis of the atom are related by
$$\mathbf{x}_n^{\text{cell}} = \mathbf{L}\mathbf{q}_n^{\text{cell}}$$
- $\mathbf{L} = (\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3) \in \mathrm{Mat}_{3,3}(\mathbb{R})$ the primitive lattice basis where coordinates are given with respect to the canonical basis.