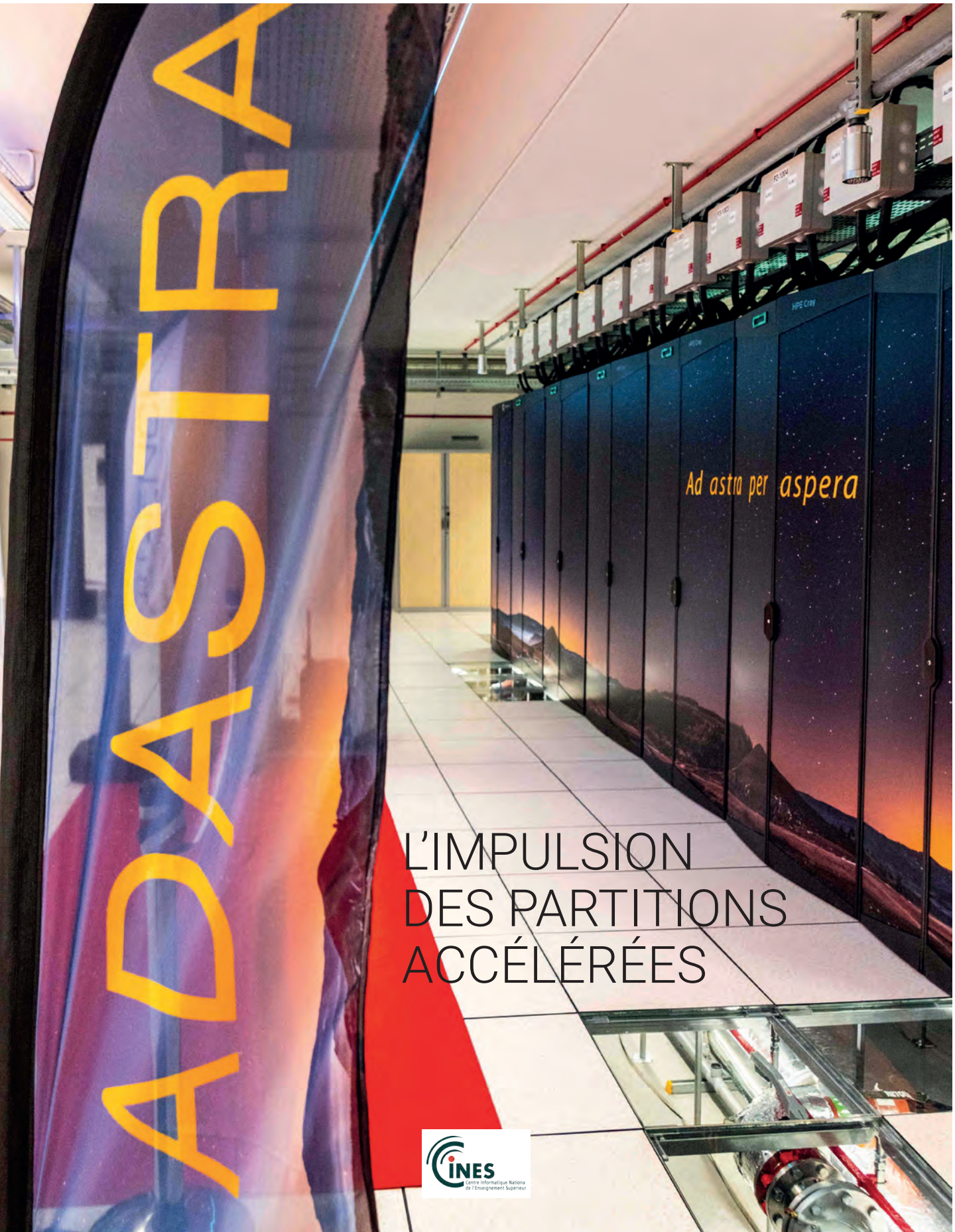
GRANDS Challenges

GENCI
Le calcul intensif au service de la connaissance

ADASTRA

Le cahier des avancées scientifiques des supercalculateurs

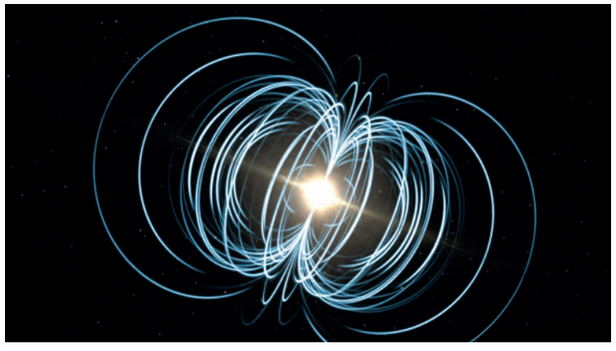
Édition février 2025

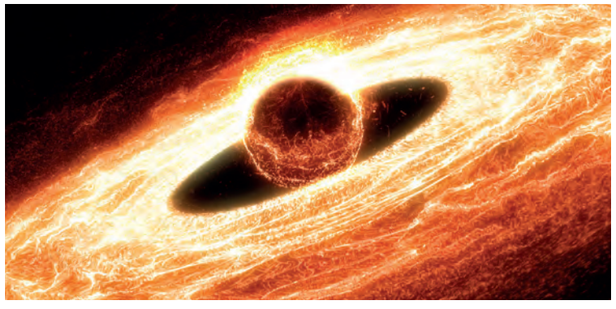


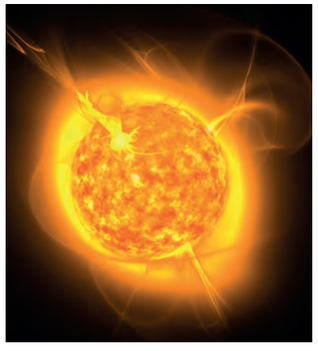
L'IMPULSION DES PARTITIONS ACCÉLÉRÉES

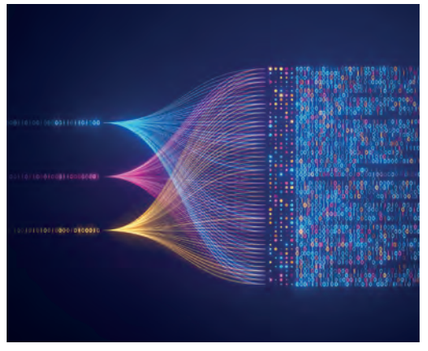
INES
Centre Informatique National de l'Enseignement Supérieur

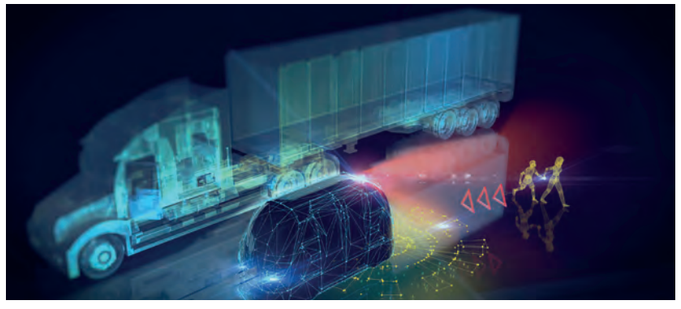
SOMMAIRE











ÉDITO

« Pour répondre aux exigences de calcul toujours plus élevées en termes de précision, de rapidité, de sécurité et de réalisme. »

Philippe LAVOCAT,
Président-directeur général de GENCI

Michel ROBERT,
Directeur du CINES (centre hébergeant et opérant le supercalculateur Adastra)

L'intelligence artificielle est devenue un enjeu de société majeur. Omniprésentes, les technologies qui lui sont rattachées constituent des chemins désormais incontournables vers l'avenir. L'« IA » est aujourd'hui présente dans presque tous les domaines de recherche. Il importe donc de renforcer considérablement les ressources françaises dans ce domaine. C'est le sens de l'acquisition des partitions accélérées d'Adastra. C'est également une étape importante de ce développement de l'IA au service de la science que marquent les Grands Challenges sur ces technologies. En effet, pour répondre aux exigences de calcul toujours plus élevées en termes de précision, de rapidité, de sécurité et de réalisme dans leur pratique de la simulation numérique, les équipes de GENCI et celle du CINES, ont souhaité mettre à la disposition des chercheurs les technologies les plus avancées. La période de mise en production de nouveaux supercalculateurs offre l'opportunité pour des utilisateurs académiques et industriels de pouvoir accéder à des ressources de calcul pouvant aller jusqu'à l'intégralité de la machine, permettant la réalisation de simulations de très grandes tailles et ainsi de mettre en œuvre des projets de simulations scientifiques hors normes, sélectionnés sur leur intérêt scientifique et sociétal. Cette période appelée « Grands Challenges », instaurée par GENCI pour toutes ses acquisitions de calculateurs depuis 2008, permet également le « rodage » des nouveaux équipements en testant leur robustesse sur des codes à grande échelle, poussant ainsi leurs capacités aux limites avant de signer la réception définitive des matériels avec l'industriel intégrateur de la machine.

Cela souligne l'importance fondamentale de la communauté des femmes et des hommes qui s'investissent avec détermination dans ces actions au service de la recherche scientifique, pour exploiter les progrès remarquables des supercalculateurs.

Pendant plus d'un an, Adastra occupe la 3e place au classement international *Green 500*. C'est le résultat des efforts entrepris, sous la houlette de GENCI, vers la recherche d'une efficacité opérationnelle énergétique maximale.

Sur les seize Grands Challenges sélectionnés sur leur ambition de dimensionnement et leur excellence scientifique, onze d'entre eux, parmi les plus aboutis, vous sont présentés dans le présent cahier. Ces projets portent sur des sujets aussi variés que l'accélération de l'occurrence de tempêtes extrêmes, le rayonnement électromagnétique des sursauts solaires en ondes « radio », ou encore l'usage d'un modèle de fondation pour la conduite autonome, etc.

UNE NOUVELLE PARTITION ACCÉLÉRÉE TRÈS INNOVANTE EN SERVICE FIN 2024

Atteindre l'excellence scientifique et concevoir des innovations ne vaut qu'à la condition de s'inscrire dans une perspective « durable ». En ce sens, le choix a été fait d'utiliser les technologies les plus efficaces énergétiquement via le calcul accéléré sur GPUs. C'est aussi ce qui explique la 3e place d'Adastra au classement international Green 500 pendant plus d'un an. Cela est le résultat des efforts entrepris par les centres de calcul nationaux sous la houlette de GENCI, notamment dans la définition des critères techniques de sélection des machines orientés dès l'origine vers la recherche d'une efficacité opérationnelle énergétique maximale : C'est la méthode de calcul du Coût Total d'Acquisition qui prend notamment en compte les coûts opérationnels dont les coûts énergétiques, dans le choix des machines.

L'ensemble de ces résultats ont été rendus possibles grâce au concept de « contrat de progrès » proposé par GENCI depuis une dizaine d'années, associant porteurs d'applications et experts du centre de calcul mais aussi ceux du vendeur (ici HPE) et ses partenaires technologies (ici AMD). Ce dispositif est un véhicule inédit pour mobiliser toutes les composantes autour du portage et de l'optimisation vers les accélérateurs ici le GPU AMD Instinct™MI250X. Il est à noter qu'une nouvelle partition accélérée utilisant l'unité de traitement accéléré (APU) très innovante AMD Instinct™ MI300A sera mise en service au dernier trimestre 2024 et permettra d'améliorer la convergence entre les charges de travail HPC et d'IA.

Cela souligne, si besoin était, l'importance fondamentale de la communauté des femmes et des hommes qui s'investissent, chaque jour, avec détermination dans ces actions au service de la recherche scientifique. Leur engagement est essentiel pour exploiter pleinement les progrès actuels remarquables des supercalculateurs et ainsi optimiser les chances de permettre des percées scientifiques remarquables voire majeures. Les directions de GENCI et du CINES s'unissent pour remercier très chaleureusement leurs équipes qui ont permis ces divers succès scientifiques. ■

Le supercalculateur

Adastra (MI250X)

Adastra 2 (MI300A)

CT1 ENVIRONNEMENT

L'ÉQUIPE
Florian PANTILLON, chargé de recherche CNRS (LAERO, Univ. de Toulouse)
Juan ESCOBAR (LAERO, Univ. de Toulouse)
Philippe WAUTELET (LAERO, Univ. de Toulouse)
Joris PIANEZZE (LAERO, Univ. de Toulouse)

Jean-Pierre CHABOUREAU (LAERO, Univ. de Toulouse)
Thibaut DAUHUT (LAERO, Univ. de Toulouse)
Christelle BARTHE (LAERO, Univ. de Toulouse)
Sophia BRUMER (LAERO, Univ. de Toulouse)

Facteur 3

Le calcul sur GPU permet d'atteindre un facteur 3 en efficacité énergétique, par rapport au CPU seul.

Accélérer la simulation des tempêtes grâce aux GPU

LE CONTEXTE. La simulation numérique de l'atmosphère joue un rôle crucial dans la compréhension et l'anticipation des événements météorologiques extrêmes. Des progrès continus en puissance de calcul ont permis d'étendre la complexité et les échelles représentées par la simulation numérique. Cependant, l'avènement d'architectures de calcul hétérogènes, combinant processeurs classiques (CPU) et graphiques (GPU), requiert l'adaptation des codes atmosphériques.

124 312 heures GPU.

Alors que les codes actuels de prévision numérique du temps calculent l'évolution de l'atmosphère à des résolutions horizontales de l'ordre de 10 km à l'échelle globale et 1 km à l'échelle régionale, le code communautaire de recherche *Méso-NH* permet de représenter des échelles et des niveaux de complexité encore inatteignables pour la prévision opérationnelle. Dans le cadre d'un projet pilote « Grand Défi » sur la partition accélérée GPU du nouveau supercalculateur Adastra, des simulations à résolution horizontale de l'ordre de 100 m ont été réalisées pour des tempêtes conduisant à des rafales de vent extrêmes.

Les simulations de tempêtes constituent la première application scientifique du portage de *Méso-NH* sur GPU. Cet exploit numérique a été rendu possible par l'inclusion de directives OpenACC dans les routines du code Fortran représentant l'essentiel du temps de calcul. Cette approche permet de calculer avec le même code sur des architectures CPU et hybrides CPU/GPU. Une attention particulière a été portée à la reproductibilité des calculs à l'octet près entre les deux types d'architecture, ce qui permet de garantir la fiabilité du portage et l'absence de bogues.

COMPRENDRE LES MÉCANISMES DE FORMATION DES RAFALES

Un point critique a été le calcul de la pression atmosphérique, qui est très complexe dans un code non-hydrostatique comme *Méso-NH* et demande l'inversion d'une équation elliptique. Un nouvel algorithme géométrique multi-grille a été implémenté pour remplacer l'approche d'origine basée sur des transformées de Fourier rapides, rendues inefficaces par les communications entre un grand nombre de processeurs GPU.
Le réalisme physique et l'efficacité numérique de *Méso-NH* sont exploités ici pour mieux comprendre les mécanismes de formation des rafales de vent à petite échelle dans les tempêtes. Les rafales sont responsables de dégâts majeurs mais restent mal comprises en raison de leur nature locale (de l'ordre de la centaine de mètres) et intermittente (de l'ordre de la seconde). Elles sont inaccessibles aux simulations atmosphériques standard et mesurées de manière très partielle par les systèmes d'observation.
Les premières simulations ont porté sur deux tempêtes représentatives des événements météorologiques extrêmes qui touchent la France : la dépression intense Alex, formée en Atlantique Nord et qui a frappé la Bretagne en octobre 2020, représentative des tempêtes hivernales, et

MÉSO-NH
Il s'agit du code communautaire français de recherche météorologique développé initialement par le LAERO et le CNRM.

ADASTRA
Ce nouveau supercalculateur au CINES, à Montpellier, est l'un des premiers au monde en termes de puissance et d'efficacité énergétique.

« Les rafales sont responsables de dégâts majeurs, mais elles restent mal comprises en raison de leur nature locale et intermittente. »

Dans le cadre du Grand Challenge, des simulatinos couplées entre le code atmosphérique Méso-NH et le code spectral de vagues WaveWatchIII ont été réalisées pour la première fois sur architecture hybride.

une ligne de grains en forme d'arc étendu qualifiée de « derecho » qui a frappé la Corse en août 2022, représentative des tempêtes méditerranéennes (voir figures).

RÉSOLUTION HECTOMÉTRIQUE
Résolution spatiale de l'ordre de 100m.

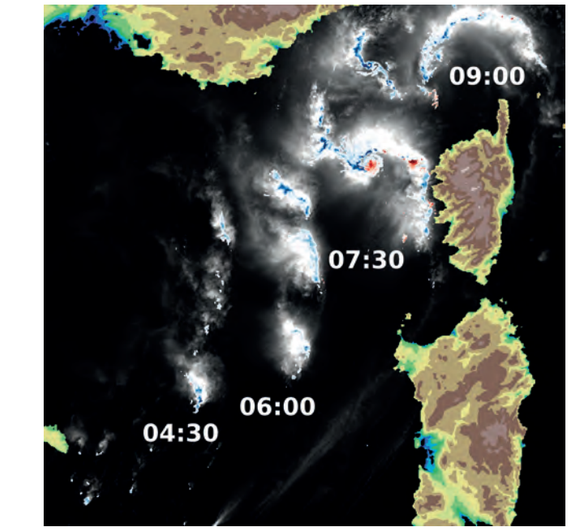
HAUTE RÉSOLUTION SPATIALE NÉCESSAIRE

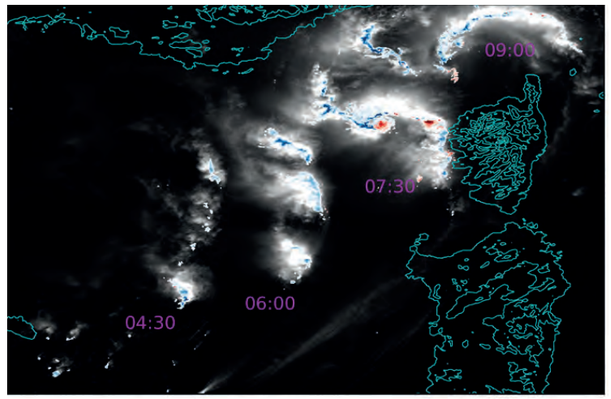
Dans chaque cas, la tempête s'étend sur plusieurs centaines de kilomètres et se propage rapidement à une vitesse de près de 100 km/h. Un domaine de simulation suffisamment étendu est nécessaire pour couvrir la période d'intensification qui dure plusieurs heures. Cependant, les processus liés à la formation de rafales se produisent à des échelles de quelques kilomètres au plus. La résolution hectométrique permet ainsi de représenter explicitement la cascade d'échelles, depuis le cœur des tempêtes jusqu'aux circulations convectives profondes et peu profondes locales à l'origine des rafales.

Ces simulations de tempêtes des latitudes tempérées ont été complétées par celles d'un système convectif organisé, observé sur la forêt amazonienne pendant la campagne de mesures *CAFE-BRAZIL* en janvier 2023, plus représentatif des tempêtes tropicales. Comme dans les deux premiers cas, la tempête s'étend sur plusieurs centaines de kilomètres mais sa propagation est plus lente en raison du plus faible vent ambient aux tropiques, ce qui requiert une simulation plus longue. Les rafales de vent associées sont relativement faibles et ne produisent pas de dégâts. Néanmoins, elles jouent un rôle crucial dans le déclenchement continuel de nouvelles cellules convectives qui maintiennent l'organisation et la propagation du système convectif. Simuler cette montée en échelle (à l'inverse de la descente en échelle décrite plus haut pour la formation des rafales) requiert également une haute résolution spatiale pour bien représenter les cellules convectives locales.

UNE PREMIÈRE SUR ARCHITECTURE HYBRIDE CPU/GPU !

En plus de la haute résolution, une simulation réaliste des tempêtes qui se développent sur mer demande une représentation détaillée des interactions entre l'océan et l'atmosphère. En particulier, la croissance et le déferlement des vagues sont contrôlés par les vents de surface mais rétroagissent aussi sur leur intensité. Dans le cadre du « Grand Défi », des simulations couplées entre le code atmosphérique *Méso-NH* et le code spectral de vagues *WaveWatchIII* ont été réalisées pour la première fois sur architecture hybride CPU/GPU, avec un surcoût numérique négligeable par rapport aux simulations atmosphériques seules. Par rapport aux codes actuels de prévision numérique du temps, de telles simulations couplées à résolution hectométrique modifient radicalement la représentation des vents sur mer comme de la hauteur des vagues, en particulier des extrêmes à l'origine de dégâts majeurs.
Pour conclure, le code *Méso-NH* est compatible avec différentes plateformes GPU-AMD et GPU-NVIDIA.
Il montre une extensibilité stable jusqu'à au moins 1024 processeurs GPU et permet d'atteindre un facteur 3 en efficacité énergétique, par rapport au calcul sur processeurs CPU seuls.
L'accélération du code communautaire *Méso-NH* sur GPU ouvre la voie à de nouvelles applications scientifiques. Ainsi, les simulations à résolution hectométrique des tempêtes permettront de mieux expliquer les processus d'accélération du vent responsables de la formation des rafales. ■





Simulation de la ligne de grains en forme d'arc étendu (derecho) du 18 août 2022 : hydrométéores nuageux (gris), vent > 100 km/h (rouge) et pluie > 50 mm/h (bleu).

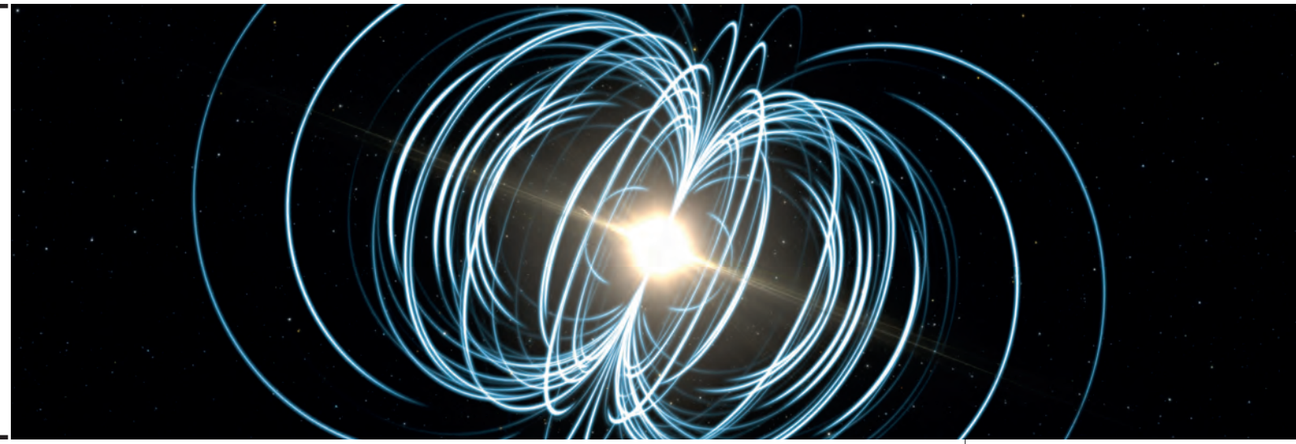
WAVEWATCHIII
Code de calcul communautaire développé par le NOAA (Agence américaine d'observation océanique et atmosphérique), dédié à la modélisation des vagues. Il inclut les dernières avancées scientifiques dans le domaine de la modélisation et de la dynamique des vagues de vent.

CT4 **GÉOPHYSIQUE ET ASTROPHYSIQUE**

L'ÉQUIPE
Rémi BOURGEOIS (CEA/DRF/MDLS)
Pascal TREMBLIN (CEA/DRF/MDLS)
Thomas PADIOLEAU (CEA/DRF/MDLS)
Samuel KOKH (CEA/DES/ISAS/DM2S/SGLS)
Yushan WANG (CEA/DRF/MDLS)
Julien BIGOT (CEA/DRF/MDLS)
Sainsbury FELIX (CEA/DRF/MDLS)
Martial MANCIP (CEA/DRF/MDLS)
Amal GUEROUDJI (CEA/DRF/MDLS)

4096^3 C'est le nombre d'éléments de la grille sur laquelle est représentée la solution, pour un total de 5To par sauvegarde.

DynoStar, simulation à grande échelle de dynamo convective stellaire



LE CONTEXTE. Ce Grand Challenge se concentre sur l'étude de la génération de champs magnétiques par la convection dite « diabatique » i.e. piloté par des termes sources thermochimiques (Tremblin et al. 2019). Ces dynamos convectives pourraient prendre part à la génération des champs magnétiques dans les étoiles chimiquement particulières. La simulation à haute résolution permise par la partition GPU d'Adastra permet de vérifier la validité de notre cadre théorique et constitue un cas idéal pour la mise en place d'outils d'analyse de données adaptés à l'ère de l'Exascale.

150 000 heures.

L'objectif principal du Grand Challenge *DynoStar* est de simuler, avec une résolution extrêmement élevée, le processus de dynamo convective diabatique dans les étoiles chimiquement particulières de type Ap et Bp. Ces étoiles se distinguent par une concentration élevée d'éléments lourds à leur surface, maintenus en place par lévitation radiative. De plus, elles possèdent un champ magnétique dont l'origine demeure difficilement explicable. Nous émettons l'hypothèse que ce champ magnétique pourrait être maintenu par un effet de dynamo convective se produisant dans les couches externes de ces étoiles. Bien que cette hypothèse semble contredite par la stabilité des atmosphères de ces étoiles aux critères de convection classiques, notre cadre théorique pour la convection diabatique i.e. avec termes sources thermochimiques dans des écoulements magnétisés, a mis en lumière de nouvelles instabilités potentielles. Celles-ci pourraient jouer un rôle crucial dans ces étoiles chimiquement particulières et être à l'origine de l'effet dynamo.

SIMULATION MHD À HAUTE RÉSOLUTION

Le deuxième aspect innovant de ce projet est notre utilisation de traitement de données in situ à l'aide de bibliothèques spécifiquement développées à la Maison de la Simulation. Les défis liés à la gestion de la quantité de données générées par les codes pré-exaflopiques sont considérables ; la quantité de données augmente avec la puissance de calcul et il devient impossible de continuer à tout sauvegarder pour analyse à posteriori. Nous avons mis en place des analyses sous forme de réductions développées indépendamment et automatiquement exécutées au moment de la production des données, ce qui évite la sauvegarde sur disque de grands volumes (5 To dans notre cas). Cette expérience met en évidence les gains critiques rendus possibles grâce à l'utilisation d'outils modernes pour la gestion de données qui deviendront de notre point de vue une nécessité avec les prochaines générations de machines.

Nous avons mené une simulation MHD à haute résolution dans un milieu stratifié, avec un champ magnétique initial faible, un gradient moyen de poids moléculaire déstabilisant et un gradient de température potentielle stabilisant. Les éléments lourds en surface induisent une instabilité convective, mais celle-ci est contrebalancée par un gradient thermique, si bien que le critère de Ledoux n'est pas satisfait. Toutefois, la présence de termes sources chimiques et thermiques permet l'instabilité

DYNAMO CONVECTIVE
La dynamo convective fait référence au processus où la convection dans un fluide conducteur, comme le fer fondu dans le noyau terrestre, génère des champs magnétiques.

TRAITEMENT IN-SITU
Le traitement in-situ désigne l'analyse et la gestion des données directement pendant que la simulation s'exécute, sans stockage intermédiaire.

« La simulation, réalisée sur 1024 GPUs, a permis de confirmer par une étude de convergence nos estimations de dynamo convective. »

Notre code implémente des méthodes modernes d'analyse numérique (relaxation et *splitting* de flux) pour une simulation stable et précise du plasma.

selon le critère diabatique (double diffusif), entraînant ainsi le mouvement convectif. Une fois les cellules de convection formées, elles induisent une intensification du champ magnétique par effet dynamo.

CONFIRMATION DES ESTIMATIONS DE DYNAMO CONVECTIVE

La simulation débute avec une résolution relativement faible de 256^3, puis est augmentée par étapes jusqu'à la résolution finale. Chaque augmentation de résolution se fait après vérification de la convergence de la solution via le calcul du spectre de puissance, comme montré dans la Figure 2. Cette stratégie permet une étude de convergence des problèmes. Notre théorie propose que la dynamo convective soit plus intense avec une faible résistivité magnétique. Nous avons vérifié cette propriété grâce à nos simulations à grande échelle. Sans résistivité physique imposée, elle est contrôlée par la diffusion numérique.

La simulation, réalisée sur 1 024 GPUs, a permis de confirmer, par une étude de convergence, nos estimations de dynamo convective. La Figure 1 montre une coupe verticale de la perturbation en densité dans le régime non linéaire, révélant une plume descendante dense et des plumes montantes légères.

5 To DE STOCKAGE POUR UNE SAUVEGARDE COMPLÈTE DE LA SIMULATION

Cette simulation a été réalisée avec le code *ARK-MHD*, utilisant les bibliothèques *Kokkos* +MPI. *Kokkos* assure le parallélisme au sein de chaque GPU et la portabilité sur diverses architectures, tandis que MPI permet l'utilisation de multiples GPUs en parallèle, assurant les communications entre eux et l'exécution synchronisée du calcul.

Notre code implémente des méthodes modernes d'analyse numérique (relaxation et *splitting* de flux) pour une simulation stable et précise du plasma. La simulation s'est déroulée sur 256 nœuds (1 024 GPUs MI250x, 130k heures) de la partition GPU du supercalculateur Adastra, atteignant une résolution maximale de 4096^3 cellules.

Une sauvegarde complète de la simulation nécessite 5 To de stockage, et la mémoire vive nécessaire pour le schéma numérique est de 35 To. Les besoins en Entrées/Sorties sont importants et multiples : moyennes globales, coupes et profils verticaux. Ces réductions sont gérées par la bibliothèque PDI (Roussel et al. 2017). PDI permet un couplage non-invasif de codes massivement parallèles avec des librairies d'Entrées/Sorties parallèles telles que hdf5. De plus, le suivi de la turbulence et le calcul du spectre de puissance ont été effectués à l'aide de la bibliothèque DEISA (Gueroudji et al. 2021) pour réaliser des transformées de Fourier en direct pendant la simulation. La stratégie adoptée pour chaque type d'Entrées/Sorties dépend de leur scalabilité. Par exemple, le calcul de moyennes globales est une opération appliquée à l'ensemble des données, donc effectué sur les GPUs utilisés par la simulation.

SIMULATIONS DE GRANDE ENVERGURE EN PERSPECTIVE

En revanche, le calcul du spectre de puissance n'implique qu'une partie mineure des données et est donc réalisé sur des ressources de calcul séparées. Cette flexibilité est permise par la bibliothèque DEISA. Nous observons la concordance entre le spectre de puissance de l'écoulement, calculé à la fois a posteriori à partir d'une tranche de données sauvegardée et in-situ, en utilisant DEISA à une résolution de 4096^3.

Ce Grand Challenge a démontré l'efficacité des bibliothèques d'Entrées/Sorties PDI et DEISA dans un contexte pré-Exascale, préparant le terrain pour des simulations de grande envergure monitorées et pilotées par des outils d'intelligence artificielle.

En effet, une piste pour le traitement des Entrées/Sorties est la détection par IA d'événements rares dans la simulation, avec une programmation de sorties à haute fréquence d'un sous-domaine centré autour de cet événement. De plus, la vérification de la convergence et le déclenchement de l'augmentation de résolution pourrait aussi être automatisé via DEISA. ■

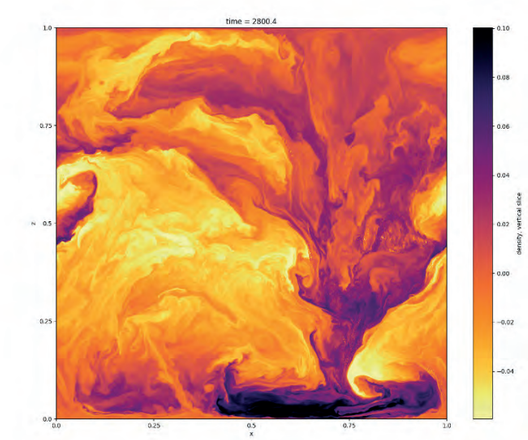


Figure 1 : Coupe verticale de la perturbation en densité dans le régime non linéaire.

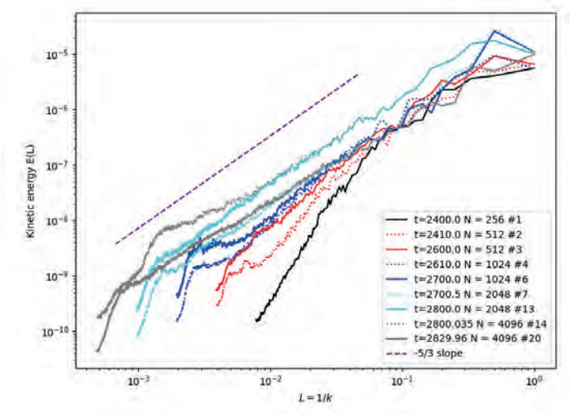


Figure 2 : Spectre de puissance de l'énergie cinétique à différents temps/résolutions.

KOKKOS
Modèle de programmation parallèle en C++ pour l'écriture d'applications portables.

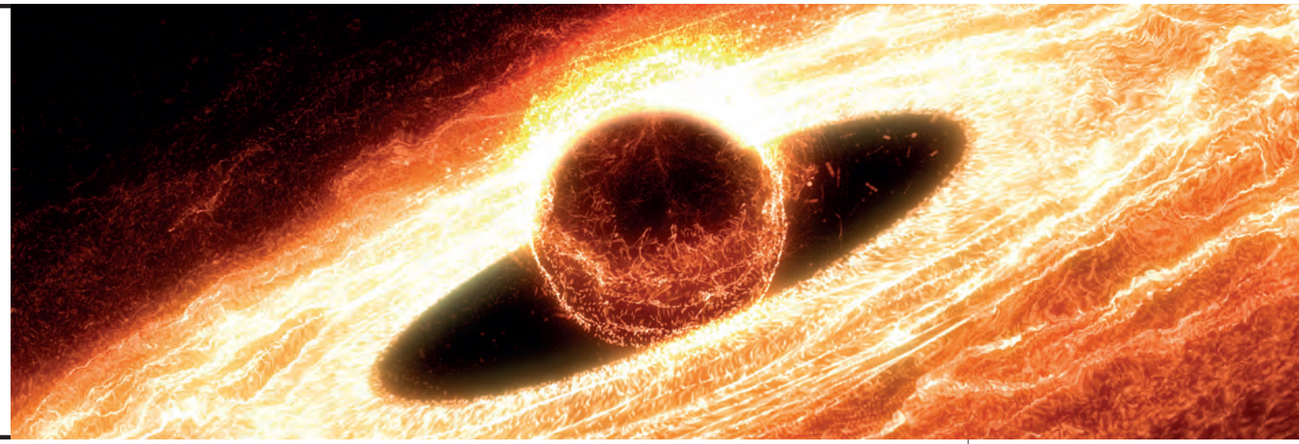
PDI
Interface pour la lecture/écriture de données massives permettant de découpler la gestion de données du calcul dans les codes de simulation».

CT4 **GÉOPHYSIQUE ET ASTROPHYSIQUE**

L'ÉQUIPE
Marc VAN DEN BOSSCHE (IPAG, CNRS, Univ. Grenoble Alpes)
Geoffroy LESUR (IPAG, CNRS, Univ. Grenoble Alpes)
Guillaume DUBUS (IPAG, CNRS, Univ. Grenoble Alpes)

1/100 C'est le rapport d'aspect du disque d'accrétion, qui est 100 fois plus fin qu'il n'est étendu radialement.

Simulations globales de disques d'accrétion à des températures extrêmes



LE CONTEXTE. Les simulations de disques d'accrétions sont nombreuses en astrophysique. Cependant la plupart des disques étudiés par des simulations sont épais, avec des rapports d'aspect d'au moins 1/10. Lorsque des disques froids se forment autours d'objets compacts, ces disques sont extrêmement fins avec des rapports d'aspect dix à cent fois moindre et les simuler demande une puissance de calcul décuplée.

300 000 heures.

Les disques d'accrétion autour d'objets compacts (naines blanches, étoiles à neutrons ou trous noirs) sont des structures extrêmement fines en raison de la forte gravité au voisinage de ces objets. Dans le meilleur des cas, c'est-à-dire quand l'objet central est une naine blanche, le rapport d'aspect de ces disques varie entre 1/100 et 1/1000 selon le régime. Quand le disque est froid, il est fin et émet un rayonnement faible, comparable à celui des étoiles des systèmes binaires dans lesquels on les trouve. Au contraire, quand il est chaud, le disque est plus épais et produit une luminosité qui peut être 100 fois plus élevée que pendant la phase froide.

Jusqu'à présent, l'étude de ces disques a porté principalement sur des disques chaud et volontairement plus épais que dans la réalité, afin de minimiser le coût numérique des simulations.

QUANTIFIER LE TRANSPORT DANS LES SYSTÈMES RÉELS

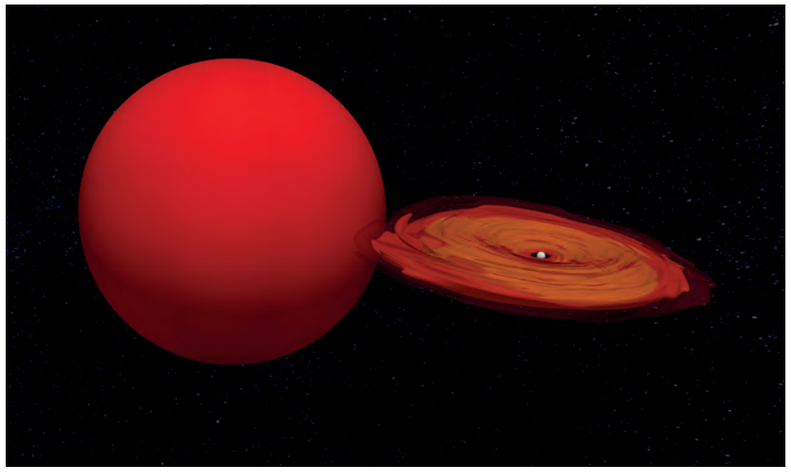
Pour ce Grand Challenge, notre objectif était d'étudier le régime froid, c'est-à-dire le moins épais, de ces disques autour de naines blanches. Ce régime correspond à la phase sombre des novæ naines, dite phase de quiescence. Plus précisément, le phénomène sur lequel nous nous sommes penchés est l'accrétion dans le disque.
Pour que de l'accrétion ait lieu, il est nécessaire que qu'une partie de la matière du disque cède son moment cinétique au reste de la matière.
Si ce n'était pas le cas, le moment cinétique total étant constant, la matière ne pourrait jamais tomber sur la naine blanche. Cette réorganisation du moment cinétique du fluide, ou transport de moment cinétique, est rendue possible par la turbulence dans le disque. Cette turbulence agit comme un effet diffusif et peut être considérée comme une viscosité effective pour le fluide.
L'origine de cette turbulence est bien établie lorsque le disque est chaud. Dans ce cas, la présence d'un champ magnétique normal au disque, même de faible amplitude, est suffisant pour qu'une instabilité magnéto-hydrodynamique produise une cascade turbulente dans le disque.
Grâce à la turbulence produite par cette instabilité, il est possible de reproduire avec des simulations, les observations de la phase chaude et brillante des novæ naines.

Cependant, pendant la phase froide, le plasma qui constitue le disque d'accrétion est peu ionisé et est peu couplé au champ magnétique extérieur. Alors,

ACCRÉTION
C'est le transport radial de matière dans l'écoulement disque. Dans le cas présent, vers la naine blanche.

« Grâce à la parallélisation de ce code GPU, il a été possible pour la première fois d'utiliser une résolution assez grande pour étudier des disques entiers avec un rapport d'aspect de 1/100. »

Grâce à ce Grand Challenge, nous avons montré que les vents magnétiques sont lancés même depuis ces disques froids et peu ionisés...



Simulation de la ligne de grains en forme d'arc étendu (derecho) du 18 août 2022 : hydrométéores nuageux (gris), vent > 100 km/h (rouge) et pluie > 50 mm/h (bleu).

RAPPORT D'ASPECT
Pour un disque, il s'agit du ratio de son épaisseur sur son rayon, cette valeur peut varier au sein du disque, avec la distance à l'objet central.

l'instabilité magnéto-rotationnelle n'est pas déclenchée, et les mécanismes usuels de transport de moment cinétique des disques chauds ne sont pas efficaces. Pourtant, grâce à des modèles basés sur des observations, il est possible de quantifier le transport dans les systèmes réels ; les valeurs ainsi obtenues pointent du doigt le fait qu'un mécanisme inconnu de transport reste efficace pendant la phase froide et peu lumineuse de quiescence.

Avec ce Grand Challenge, et grâce au nouveau code accéléré *Idéfix*, nous avons pu réaliser des simulations de tels disques dans des régimes de température, c'est-à-dire d'épaisseur du disque, réalistes.
Grâce à la parallélisation de ce code GPU, il a été possible pour la première fois d'utiliser une résolution assez grande pour étudier des disques entiers avec un rapport d'aspect de 1/100, c'est-à-dire dix fois plus mince que les études globales précédentes.
Afin de modéliser la phase froide, nous avons inclus une résistivité pour le plasma, pour rendre compte du faible couplage au champ magnétique.

UN VENT MAGNÉTIQUE DE GRANDE ÉCHELLE !

Les simulations réalisées dans le cadre de ce Grand Challenge ont montré que le mécanisme inconnu permettant le transport de moment cinétique dans un disque froid, fin et peu ionisé était un vent magnétique de grande échelle. Ce mécanisme avait déjà été suggéré par des simulations zoomées, à plus petite échelle, mais sa validité pour l'entièreté du disque restait spéculative.

Le lancement d'un tel vent magnétique peut être compris comme la réaction du champ magnétique vertical à grande échelle à une accrétion initiale. Dans les régions où le couplage entre le plasma et le champ est efficace, l'écoulement du disque d'accrétion va tordre les lignes de champ magnétique. En réaction, le champ magnétique impose une force au plasma, la force de Lorentz, qui va donner lieu à l'éjection verticale d'une fraction de la matière du disque, accompagnée de couples aux surfaces du disque qui permettent le transport de moment cinétique. Dans le cas des disques chauds, c'est dans le disque que le couplage entre plasma et champ magnétique est efficace, et dans ces disques on a alors un vent magnétique et de la turbulence. Dans les disques froids étudiés ici, l'intérieur du disque est peu turbulent, mais à ses surfaces le plasma y est moins dense et plus chaud.

Là, l'instabilité magnéto-rotationnelle peut se développer et un vent peut être lancé. Si le transport de moment cinétique dû à la turbulence est faible, celui lié aux couples du vent magnétique est significatif, même dans un disque globalement froid et peu ionisé.

UN CHANGEMENT DE PLAN DU DISQUE JAMAIS OBSERVÉ

À notre grande surprise, après une demi-douzaine d'orbites du système binaire, nous avons observé dans toutes les simulations avec champ magnétique un changement considérable de géométrie. Le disque, initialement dans le plan du système binaire s'incline rapidement dans une direction arbitraire sur une échelle de temps liée à la vitesse du son. L'orientation de l'inclinaison du disque tourne alors à une fréquence de quelques pourcents de celle du système binaire. Cette configuration avec un disque incliné avait déjà été proposée pour expliquer des épi-éruptions observées lors de la phase chaude de certains systèmes. Cependant, jamais auparavant un tel changement de plan du disque n'avait été observé dans de telles simulations, avec un système binaire dans lequel les deux étoiles ont des spins parallèles et le champ magnétique est initialement symétrique. Au contraire, obtenir des disques inclinés dans de telles configurations n'avait jamais été réussi jusqu'à présent.

Ainsi, grâce à ce Grand Challenge, nous avons montré que les vents magnétiques sont lancés même depuis ces disques froids et peu ionisés et contribuent de manière significative à l'accrétion et au transport de moment cinétique même pendant la phase de quiescence. Nous avons aussi obtenu des disques fortement inclinés, qui sont des candidats potentiels pour expliquer des sursauts de luminosités périodiques dans certains systèmes. ■

L'INSTABILITÉ MAGNÉTO-ROTATIONELLE
Mécanisme magnéto-hydrodynamique qui couple les particules fluides sur une même ligne de champ magnétique, comme si elles étaient reliées par un ressort, et permet un échange de moment cinétique.

CT4 **GÉOPHYSIQUE ET ASTROPHYSIQUE**

L'ÉQUIPE
Vadim MONTEILLER (Laboratoire de Mécanique et d'Acoustique)

10 Hz

Inversion de formes d'ondes complètes (FWI) à 10 Hz en milieu viscoélastique.

Imagerie sismique haute résolution en milieu viscoélastique

LE CONTEXTE. Les méthodes d'imagerie sismique révèlent les structures géologiques sous la surface terrestre, cruciales pour la géodynamique, le risque sismique et la gestion des ressources naturelles. Utilisant le code de calcul de propagation d'ondes sismiques *SPECFEM3D*, nous avons inversé un jeu de données issues d'une campagne sismique en mer du Nord. La puissance de calcul disponible sur ADASTRA nous a permis d'améliorer la résolution des images.

400 000 heures.

Les méthodes d'imagerie sismique, ou sismologiques, jouent un rôle essentiel dans l'exploration des structures géologiques sous la surface terrestre. Ces techniques fournissent des images détaillées du sous-sol, cruciales pour comprendre les processus géodynamiques, évaluer le risque sismique et gérer les ressources naturelles.
Cependant, obtenir des images de haute résolution nécessite des capacités de calcul considérables et l'utilisation de codes massivement parallèles et optimisés. Les images sismiques permettent de visualiser les différentes couches géologiques et d'identifier des structures spécifiques comme des réservoirs d'hydrocarbures, des failles sismiques ou des formations minérales. Ces informations sont indispensables pour :

- **La géodynamique :** Comprendre les mouvements des plaques tectoniques et les processus internes de la Terre.
- **Le risque sismique :** Aide à la cartographie des zones à risque.
- **La gestion des ressources naturelles :** Localiser et exploiter efficacement les ressources en hydrocarbures ou en minéraux.

TROIS SIMULATIONS PAR SOURCE SISMIQUE

L'inversion de formes d'ondes complètes (FWI, *Full Waveform Inversion*) permet d'obtenir des images du sous-sol en comparant les ondes sismiques enregistrées à la surface avec les ondes modélisées. On peut, ainsi, ajuster progressivement le modèle du sous-sol pour améliorer la concordance entre les données et les sismogrammes modélisés par optimisation numérique.
La méthode utilisée est basée sur le calcul du gradient de la fonction coût par l'état adjoint. Ainsi, chaque itération nécessite 3 simulations par source sismique.
Un processus complet d'imagerie peut nécessiter plusieurs dizaine ou centaines de millier de simulations. Ce qui rend crucial l'utilisation d'un solveur efficace et optimisé pour les plateformes de calcul disponibles.
Le code *SPECFEM3D* permet de résoudre l'équation de l'élastodynamique pour des milieux 3D viscoélastique et anisotropes par la méthode des éléments finis spectraux. Ce code est optimisé pour le calcul sur GPU et ainsi très efficace pour les simulations d'ondes sismiques à haute résolution.
Grâce à l'allocation de 400 000 heures de calcul GPU, nous avons pu mettre en œuvre la méthode FWI en milieu viscoélastique

INVERSION DE FORMES D'ONDES COMPLÈTES (FWI)
Méthode d'imagerie sismique du sous-sol par ajustement des sismogrammes enregistrées et modélisées.

« La puissance de calcul des GPU permet désormais d'imager la croûte terrestre et le manteau supérieur à des résolution sans précédent. »

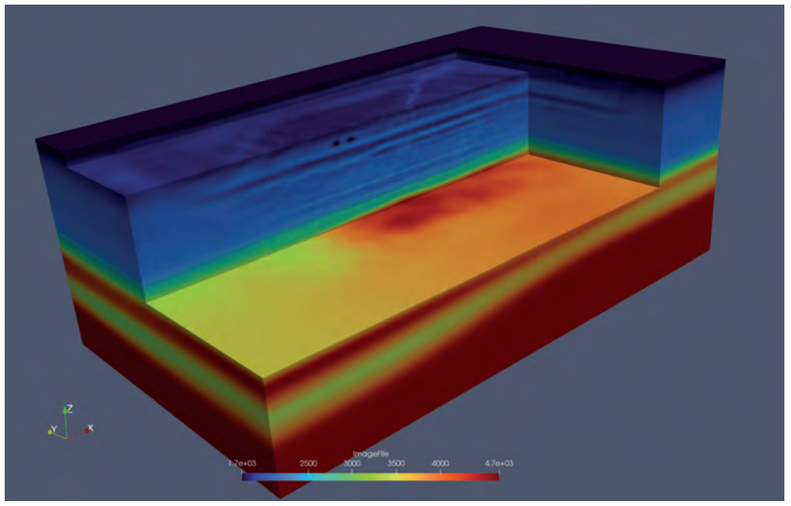


Image de la vitesse des ondes de compression dans le modèle obtenu.

Jusqu'à 100 nœuds ADASTRA ont été utilisés. Ce qui correspond à 400 GPU MI250x d'AMD avec 800 processus MPI.

anisotrope sur un cas d'étude d'un réservoir d'hydrocarbure en mer du Nord.
Une zone s'étendant sur 16 x 10 km a été instrumentée par des OBS et des tirs ont permis de générer les ondes sismiques. Nous avons utilisé les données 3 composantes enregistrées par une nappe de 128 OBS placées à environ 100 mètres de fond.

OCEAN BOTTOM SEISMOMETER (OBS)
Instrument permettant de mesurer les ondes sismiques sur le fond de la mer.

500 FOIS PLUS DE PUISSANCE DE CALCUL

Quelques 70 000 sources sismiques ont été déclenchées à 5 mètres de profondeur avec un espacement de 25 mètres. Un total de 26 millions de traces sismiques de 10 secondes ont été inversées. Le volume de stockage nécessaire pour ces données est d'environ 1 To.
La modélisation considérée est un milieu viscoélastique anisotrope qui représente mieux la physique de la propagation des ondes que l'approche classique qui considère que la Terre se comporte comme un milieu fluide et donc néglige les ondes de cisaillement et les différentes conversions de phases. En général, cette approximation acoustique est faite pour réduire le coût de calcul.
Le passage de l'approximation acoustique au milieu viscoélastique nécessite de réduire la taille de la maille de discrétisation d'un facteur 4, due principalement à la faible vitesse des ondes de cisaillement dans les sédiments. Il faut ainsi considérer des maillages 64 fois plus volumineux, en plus des champs à trois composantes. Ainsi, il faut 192 fois plus de capacité mémoire sur le cas élastique qu'avec l'approximation acoustique. Mais en tenant compte de la complexité plus importante du schéma numérique dans le cas élastique, on peut considérer que passer de l'approximation acoustique au cas élastique demande 500 fois plus de puissance de calcul.

RÉALISER UNE IMAGE DES PROPRIÉTÉS PHYSIQUES DU SOUS-SOL

La capacité de calcul disponible sur la partition GPU de ADASTRA permet de réaliser les calculs dans des milieux viscoélastiques.
Principalement due à l'accélération, d'un à deux ordres de grandeurs, du code *SPECFEM3D* obtenue sur GPU par rapport au calcul sur CPU. De plus la capacité à faire fonctionner simultanément plusieurs centaines de GPU permet de distribuer le travail efficacement grâce aux communications asynchrones de la bibliothèque MPI.
Au final, les capacités de calculs disponibles sur ADASTRA nous ont permis de réaliser une image des propriétés physiques du sous-sol de la zone considérée. Nous avons pu mener les calculs jusqu'à 10 Hz pour obtenir l'image finale la mieux résolue.
Cela a nécessité de réaliser environ 100 000 simulations dans des maillages pouvant contenir jusqu'au million d'éléments, ce qui correspond à environ 350 millions de degrés de liberté. Nous avons utilisé jusqu'à 100 nœuds ADASTRA, ce qui correspond à 400 GPU MI250x d'AMD avec 800 processus MPI.

IMAGER LA CROÛTE TERRESTRE ET LE MANTEAU SUPÉRIEUR ?

L'accès aux supercalculateurs GPU tels que ADASTRA permet de relever le défi numérique que représente l'inversion de formes d'ondes complètes dans des milieux viscoélastiques 3D.

Le grand défi GENCI dont nous avons bénéficié sur ADASTRA a démontré la faisabilité de cette approche sur un cas d'étude issue d'une campagne d'acquisition sismique.
Cela ouvre des perspectives pour des études à plus grandes échelles. Comme, par exemple, des études géodynamiques plus fondamentales, dans des zones des subductions ou des orogènes, afin d'imager la croûte terrestre et le manteau supérieur à des résolutions jamais atteintes jusqu'à présent. ■

DEGRÉS DE LIBERTÉ (DDL)
Désigne le nombre de variables aléatoires qui ne peuvent être déterminées ou fixées par une équation.

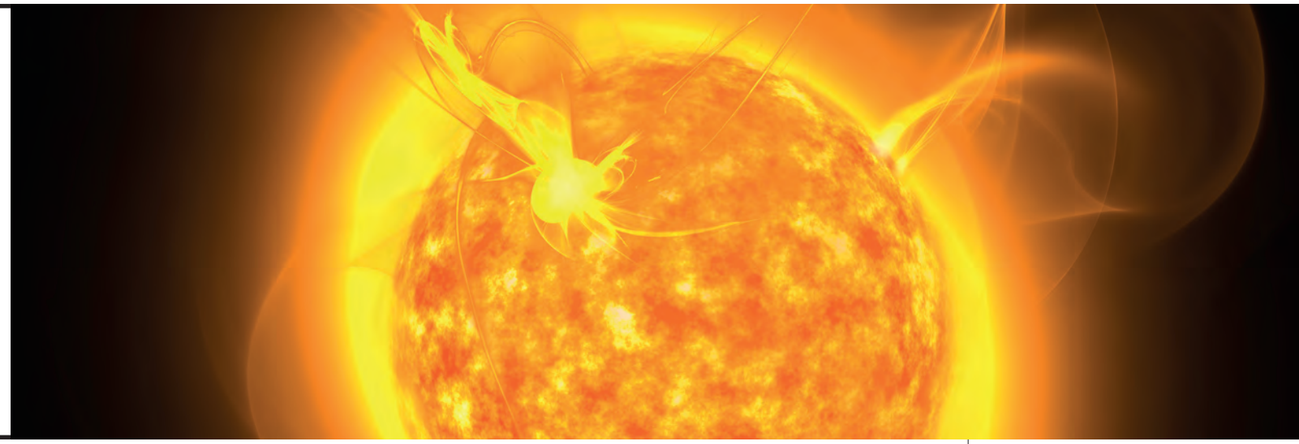
CT5 **PHYSIQUE THÉORIQUE & PHYSIQUE DES PLASMAS**

L'ÉQUIPE
Arnaud BECK (Lab. Leprince-Ringuet, CNRS/IPP)
Catherine KRAFFT (Lab. Physique des Plasmas, CNR/UPSAy/SU/IPP)
Philippe SAVOINI (Lab. Physique des Plasmas, CNR/UPSAy/SU/IPP)
Alexander VOLOKITIN (Lab. Physique des Plasmas, CNR/UPSAy/SU/IPP)

Francisco Javier POLANCO-RODRIGUEZ ((Lab. Physique des Plasmas, CNR/UPSAy/SU/IPP)
Charles PROUVEUR (Maison de la Simulation, CNRS/CEA/UPSAy)

x 10 C'est environ le gain en vitesse d'exécution lors du passage de 4 096 cœurs CPU de type AMD Rome (Epyc) à 16 MI 250 (128 GPU).

Comprendre le rayonnement électromagnétique des sursauts radio solaires



LE CONTEXTE. Parmi les émissions les plus intenses du système solaire, les sursauts radio solaires sont produits par des faisceaux d'électrons accélérés dans la couronne solaire lors d'éruptions. Lors de leur propagation dans le milieu interplanétaire, ceux-ci excitent une turbulence d'ondes électrostatiques qui émet à son tour des ondes électromagnétiques via une chaîne complexe de processus qui restent encore à élucider. Les simulations cinétiques haute-performance de la dynamique faisceau-plasma permettent maintenant des études réalistes de ces phénomènes.

200 000 heures GPU.

Lors des éruptions solaires et des éjections de masse coronale, des sursauts radio sont produits par des faisceaux d'électrons accélérés dans la couronne solaire. Ceux-ci génèrent une turbulence d'ondes électrostatiques (ES) émettant à son tour des ondes électromagnétiques (EM) à la fréquence plasma électronique et ses harmoniques, via une chaîne complexe de mécanismes successifs où les interactions entre ondes, faisceaux d'électrons et plasmas solaires jouent un rôle majeur. L'analyse des émissions EM permet d'obtenir des informations clefs sur les phénomènes d'accélération et de transport des électrons énergétiques ainsi que sur les mécanismes de dissipation, de conversion et de transfert d'énergie aux différentes échelles (turbulence) dans le vent et la couronne solaires, ouvrant ainsi des perspectives majeures pour la compréhension des phénomènes physiques à l'œuvre dans les atmosphères solaire et stellaires.

Aujourd'hui, des radiotélescopes terrestres de nouvelle génération, comme *LOFAR* et *NenuFAR* , observent le Soleil avec des résolutions temporelle et spectrale sans précédent et une sensibilité remarquable, tandis que les récentes missions *Parker Solar Probe* et *Solar Orbiter* explorent l'atmosphère solaire à des distances minimales du Soleil jamais atteintes. En outre, des expériences d'interaction laser-plasma ont récemment modélisé en laboratoire des conditions et des processus relevant des sursauts solaires.

UN DÉFI MAJEUR RELEVÉ SUR ADASTRA GPU

Bien que les sursauts radio soient observés depuis des décennies et que les travaux initiés dans les années 50 pour explorer les mécanismes d'émission d'ondes EM aient été constants, des questions majeures restent irrésolues.

En particulier, des observations spatiales in-situ ont montré que l'atmosphère solaire présente des fluctuations aléatoires de densité de petites échelles, dues à des phénomènes de turbulence, d'un niveau moyen de quelques pourcents de la densité moyenne des plasmas, affectant fortement la turbulence d'ondes ES générée par les faisceaux d'électrons et les processus qui les transforment en ondes EM. Ces circonstances remettent en question les théories existantes, développées pour des plasmas homogènes, ainsi que la quasi-totalité des simulations numériques effectuées à ce jour sur le sujet.

LOFAR & NENUFAR
LOFAR (LOw Frequency ARray) est un interféromètre constitué de plus de 50 000 antennes en Europe, dont 60 % aux Pays-Bas, réparties en 52 stations. NENUFAR (New Extension in Nançay Upgrading LOFAR) est un radiotélescope installé sur le site de la station de radioastronomie de Nançay en Sologne, observant dans des basses fréquences de l'ordre de 10 MHz et 85 MHz.

« L'accélération des calculs grâce aux GPU permet d'expliquer pour la première fois la dynamique d'un processus physique essentiel. »

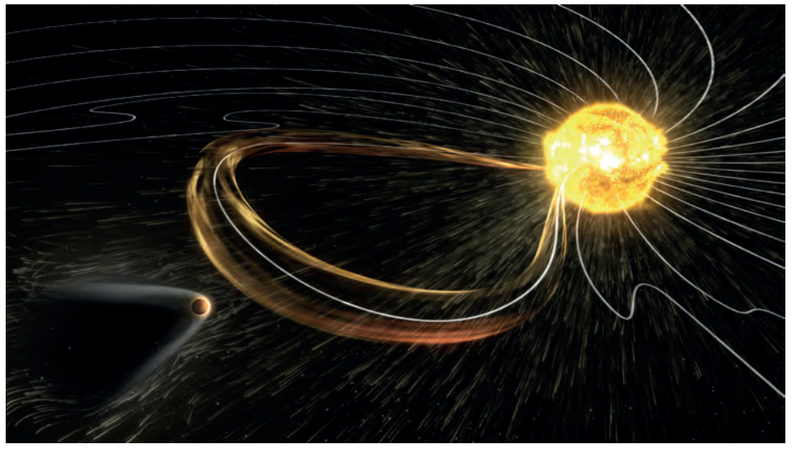


Illustration représentant une éruption solaire avec éjection de masse coronale dans le milieu interplanétaire
(Crédit : NASA's Scientific Visualization Studio)

Toutes ces modifications permettent des simulations beaucoup plus rapides et par conséquent des études portant sur un large éventail de paramètres.

Ainsi, l'un des problèmes essentiels à résoudre consiste (i) à identifier et caractériser les mécanismes de conversion de l'énergie ES de la turbulence d'ondes dans la source de plasma coronal en énergie EM rayonnée à des fréquences spécifiques dans le vent solaire, et (ii) à déterminer comment le transport de l'énergie EM s'effectue hors de cette source, sur de longues distances à travers la couronne et le vent solaires, caractérisés par des inhomogénéités de densité à la fois de grandes et de petites échelles.
La nécessité d'élaborer des approches théoriques innovantes ainsi que des simulations numériques haute-performance capables de décrire la physique des sursauts dans des plasmas solaires réalistes constitue un défi majeur qui a pu être en partie relevé grâce au Grand Challenge sur Adastra GPU.

RÉALISTE
Notre groupe est actuellement le seul au monde dans notre domaine de recherche à réaliser des simulations aussi avancées.

MACRO-PARTICULES
Une simulation typique utilisant 128 GPU permet de simuler la dynamique d'environ 500 millions de particules par seconde.

SUIVRE LA DYNAMIQUE RAPIDE DES ÉLECTRONS ET LA DYNAMIQUE LENTE DES IONS

Dans cette optique, des simulations bi-dimensionnelles massivement parallèles ont été réalisées sur la partition GPU MI 250 de la machine Adastra à l'aide du code *particle-in-cell* (PIC) *SMILEI*.
Celui-ci est utilisé pour modéliser, au niveau cinétique, les processus d'interaction ondes-ondes, ondes-particules et ondes-plasma à l'œuvre dans les plasmas du vent et de la couronne solaires, qui sont à l'origine des émissions EM. Pour atteindre une description suffisamment réaliste des phénomènes étudiés, des contraintes numériques très exigeantes doivent être satisfaites. L'inclusion de fluctuations de densité aléatoires inhérentes à ces milieux nécessite l'utilisation d'un nombre très important de macro-particules par cellule afin de réduire le bruit statistique inhérent aux codes particulaires.

D'autre part, les processus à étudier incluent à fois des ondes ES et EM, dont les longueurs d'ondes s'étendent sur deux ordres de grandeur, nécessitant des boîtes de simulations de grande taille, sans compter que seule une résolution spatiale élevée permet de discriminer les différents modes EM lorsque la magnétisation du plasma est faible, comme celle du vent solaire.
Finalement, les ondes participant aux mécanismes d'émission EM sont à la fois de basses et de hautes fréquences, si bien qu'il faut suivre à la fois la dynamique rapide des électrons et la dynamique lente des ions, nécessitant des simulations très longues.

NOUVEAUX RÉSULTATS SUR LES MÉCANISMES DE GÉNÉRATION D'ONDES EM

Pour atteindre le niveau de performance nécessaire à ce Grand Challenge, le code a été porté sur les GPU AMD en utilisant principalement les librairies *Thrust* et *openMP*. Plusieurs algorithmes au cœur de *SMILEI*, comme par exemple le tri des particules, ont dû être repensés pour profiter au mieux des performances du GPU. L'opération dite de projection, qui est la plus coûteuse, a été réimplémentée en utilisant l'API spécifique - ici HIP - au GPU visé. Cet effort a permis des optimisations plus bas niveau. Enfin, des communications MPI GPU-aware ont été mises en place autant que possible.

Toutes ces modifications permettent des simulations beaucoup plus rapides et par conséquent des études portant sur un large éventail de paramètres, décrivant un panorama physique étendu des conditions in-situ observées par les satellites.
Elles ont permis l'obtention de nouveaux résultats sur les divers mécanismes de génération d'ondes EM rayonnées à la fréquence plasma électronique ainsi que sur l'influence des caractéristiques physiques du plasma solaire (magnétisation, températures des ions et des électrons, niveau moyen des fluctuations de densité) sur ces processus [Krafft et Savoini, 2023, 2024a, Krafft et al., 2024b]. ■

Krafft, C. & Savoini P. (2023), Dynamics of Two-dimensional Type III Electron Beams in Randomly Inhomogeneous Solar Wind Plasmas, The Astrophysical Journal, 949:24, doi:10.3847/1538-4357/acc1e4
Krafft, C. & Savoini P. (2024a), Electrostatic Wave Decay in the Randomly Inhomogeneous Solar Wind. Astrophysical Journal Letters, 964L30, doi:10.3847/2041-8213/ad3449
Krafft, C., Savoini P. & Polanco-Rodriguez, F.J. (2024b), Mechanisms of Fundamental Electromagnetic Wave Radiation in the Solar Wind, Astrophysical Journal Letters, 967:L20, doi: 10.3847/2041-8213/ad47b5

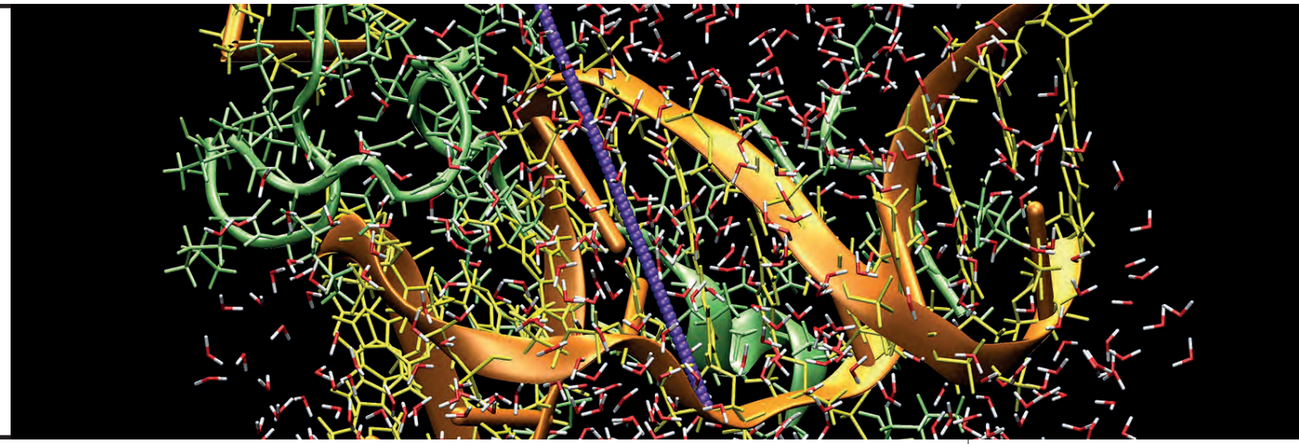
HIP
***Heterogeneous-compute Interface for Portability* est un environnement de programmation C++ proposé par AMD pour permettre la translation automatique de codes en CUDA vers un format portable utilisable à la fois sur environnements AMD et NVIDIA.**

CT7 **MODÉLISATION MOLÉCULAIRE APPLIQUÉE À LA BIOLOGIE**

L'ÉQUIPE
Aurélien DE LA LANDE (CNRS, Univ. Paris Saclay)
Karim HASNAOUI (CNRS, IDRIS, Maison de la Simulation)

40 x C'est le gain maximal de temps d'exécution de l'étape propagation lors du passage de 64 CPU à 8 GPU sur Adastra.

Irradiation ionisante de la matière biologique



LE CONTEXTE. La simulation ab initio de la dynamique électronique induite par l'irradiation ionisante de la matière est précieuse pour aider à comprendre les mécanismes moléculaires à la base de l'efficacité des radiothérapies du XXIe siècle. Grâce au Grand Challenge, l'équipe de l'Institut de Chimie Physique a étudié le dépôt d'énergie de protons dans l'eau et dans le nucléosome, un complexe d'ADN et de protéines au cœur des chromosomes.

88 000 heures.

Les rayonnements ionisants tels que les photons X ou gamma ou les noyaux d'atomes rapides constituent une classe particulière de rayonnements.
Ils ont des origines naturelles (rayons cosmiques, rochers, mers …) ou anthropiques (médecine, installations nucléaires). Les rayonnements ionisants jouent un rôle important dans divers contextes. En médecine, ils sont par exemple à la base de stratégies de défense puissantes pour traiter les cancers. L'hadronthérapie en particulier est une radiothérapie prometteuse pour éviter les effets collatéraux.

Un autre exemple est celui de l'industrie aérospatiale. Les engins électroniques ou les humains envoyés au-delà de la magnétosphère terrestre sont en effet sujets à être irradiés par le vent solaire et les rayons cosmiques avec des conséquences techniques et médicales délétères.

Les rayonnements ionisants interagissent si fortement avec la matière qu'ils en expulsent des électrons, laissant derrière eux de la matière ionisée et porteuse de forts excès d'énergie. En se dissipant dans le milieu, l'énergie déposée induit de profondes modifications chimiques délétère dans certains cas (endommagements) ou bénéfique pour d'autres (radiothérapies). Ces mécanismes premiers sont encore très mal connus et constituent un axe de recherche dynamique de l'Institut de Chimie Physique (ICP).

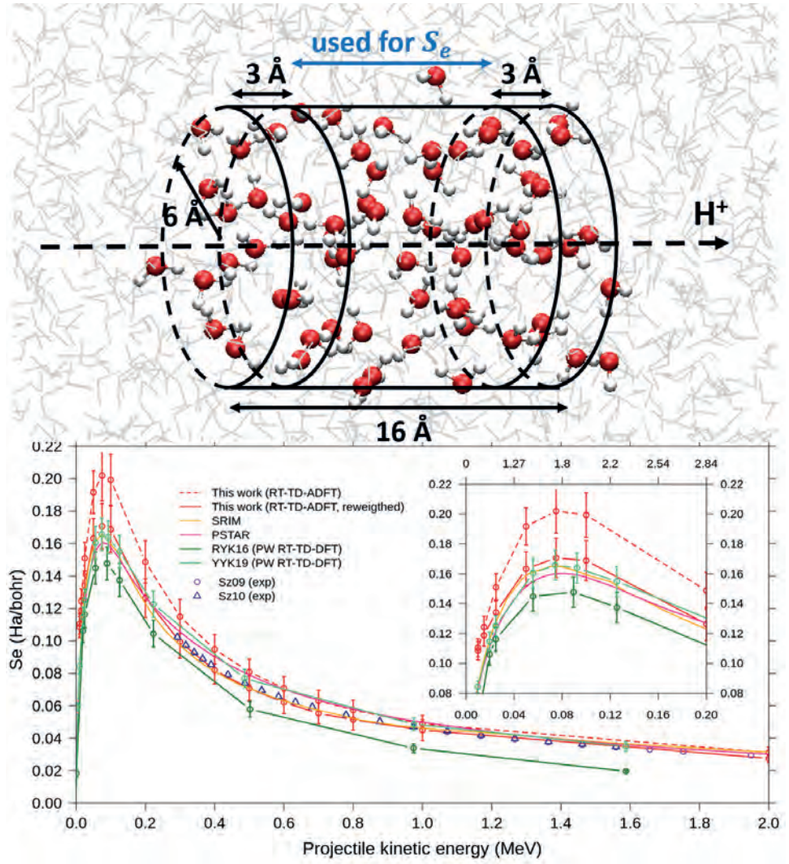
UNE GRANDE PUISSANCE DE CALCUL NÉCESSAIRE

Les simulations ab initio ont un rôle majeur à jouer pour comprendre à l'échelle moléculaire les mécanismes d'ionisation, de dépôt d'énergie et réactivité chimique radioinduites. Pour y parvenir, l'équipe *ThéoSim* de l'ICP développe un code de simulation dédié reposant sur la théorie de la fonctionnelle de la densité dépendant du temps (RT-TD-DFT en anglais) [1].
L'approche consiste à simuler pas-à-pas, le mouvement des électrons de la matière au passage d'un photon de haute énergie ou d'un ion rapide. Ce faisant, l'énergie déposée lors de l'interaction peut être estimée tandis qu'une analyse fine des états excités du nuage électronique peut être faite. Les simulations RT-TD-DFT nécessitent une grande puissance de calcul et des algorithmes hautement performants. Peu avant l'ouverture du Grand Challenge

DFT
Approche de mécanique quantique pour décrire et comprendre la physique des électrons de la matière à l'échelle microscopique.

« Pour la première fois des simulations ab initio ont estimé un pouvoir d'arrêt électronique sur un système biologique aussi complexe. »

la simulation reproduisait de manière satisfaisante le pouvoir d'arrêt électronique sur un large gamme d'énergie cinétique du proton.



Haut : *un modèle moléculaire pour estimer le pouvoir d'arrêt électronique d'un proton (H+) pénétrant de l'eau.*

Bas : *validation de l'approche théorique par comparaison à des données expérimentales.*

LE POUVOIR D'ARRÊT
Une quantité clé à connaître pour ajuster l'énergie du rayonnement en fonction de la profondeur des tumeurs dans les tissus.

Adastra, le logiciel *deMon2k* avait été porté sur GPU en partenariat avec Karim Hasnaoui de l'IDRIS et la Maison de la Simulation. L'arrivée d'Adastra a permis de tirer pleinement partie des capacités du code. Des profils de performances ont été réalisés sur des agrégats d'eau et des fullerènes (C_{20} à C_{540}) montrant une accélération de 15 à 40 lors du passage CPU à GPU en fonction de la taille des systèmes [2].

DÉFI RELEVÉ GRÂCE AUX MOYENS DU GRAND CHALLENGE

Une première étape a été la validation de l'approche par comparaison à l'expérience. À cette fin, nous avons considéré le pouvoir d'arrêt électronique, défini comme la quantité d'énergie cinétique perdue par le rayonnement par unité de déplacement. Le pouvoir d'arrêt traduit donc la capacité de la matière à freiner le rayonnement. Coté simulation, un paramètre important est la base d'orbitales atomiques utilisée pour exprimer mathématiquement les fonctions d'ondes décrivant les électrons. Dans le cas de l'interaction avec le rayonnement ionisant, ces bases doivent décrire précisément et les électrons liés aux atomes et les électrons émis dans le continuum lors de l'irradiation. Il s'agit d'un défi que les moyens du Grand Challenge nous ont permis de relever.
Les courbes de pouvoir d'arrêt de proton pénétrant l'eau liquide ont ainsi été calculées avec sur le banc d'essai un large panel de bases d'orbitales atomiques développées pour l'occasion.

SIMULATION DE L'IRRADIATION DU NUCLÉOSOME

Nous avons montré que la simulation reproduisait de manière satisfaisante le pouvoir d'arrêt électronique sur un large gamme d'énergie cinétique du proton (10 keV à 1 MeV) [3]. Les simulations accumulées lors du Grand Challenge ont, en outre, permis de nourrir un algorithme d'apprentissage d'un réseau de neurones pour, à l'avenir, remplacer les simulations RT-TD-DFT et converger statistiquement les résultats (thèse de doctorat, Damien Tolu, Université Paris Saclay, 2023).
Forts de ce succès nous avons entrepris de simuler l'irradiation du nucléosome, lequel est un complexe d'ADN et de protéines qui constituent l'unité de base de la chromatine. À ce jour nous ne connaissons pas le pouvoir d'arrêt qu'offre cette structure au passage d'ions rapides. Pour la première fois des simulations ab initio ont estimé un pouvoir d'arrêt électronique sur un système biologique aussi complexe. Des premières simulations ont été réalisées dans le cadre du grand Challenge [2].

UNE RÉFÉRENCE POUR DE FUTURES ÉTUDES D'IRRADIATION ?

En conclusion, le Grand Challenge Adastra nous a permis de faire un saut en avant dans la modélisation ab initio de l'irradiation ionisante de la matière.
Nous pensons que l'article découlant de ce travail [3] fournira une référence pour de futures études d'irradiation de systèmes variés.
À l'ICP, de nouvelles campagnes de simulations se poursuivent sur Adastra dans le cadre des appels du GENCI (sujet de thèse de Th. Boukéké-Lesplulier). ■

[1] K. A. Omar et al. Eur. J. Phys. S.T. 2023, 223, 2167.
[2] P. A. Martinez et al. Comput. Phys. Commu. 2024, 295, 108946. [3] R. Tandiana, J. Chem. Theor. Comput. 2024, 19, 7740.

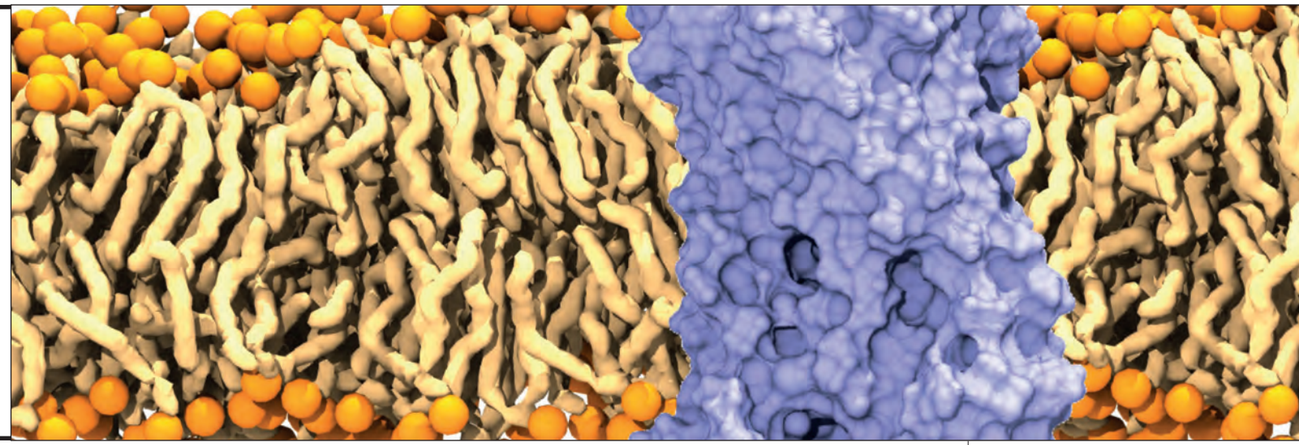
NUCLÉOSOME
Complexe moléculaire de protéines et d'ADN à la base du stockage de l'information génétique et de l'expression des gènes.

CT7 **MODÉLISATION MOLÉCULAIRE APPLIQUÉE À LA BIOLOGIE**

L'ÉQUIPE
Nicolas FLOQUET (IBMM CNRS UMR5247, Univ. de Montpellier, ENSCM)
Maxime LOUET (IBMM CNRS UMR5247, Univ. de Montpellier, ENSCM)
Rita ROESSNER (IBMM CNRS UMR5247, Univ. de Montpellier, ENSCM)
Hafez RAZMAZMA (IBMM CNRS UMR5247, Univ. de Montpellier, ENSCM)

2000 µs
C'est le temps de simulation «tout-atomes» généré pour ce projet.

Exploration du paysage conformationnel du récepteur de la Ghréline



LE CONTEXTE. Ces dernières années, des centaines de structures expérimentales ont été décrites décrivant les interactions entre récepteurs couplés aux protéines G (RCPG) et différentes protéines partenaires ou petites molécules (ligands). Ces structures ont ouvert de nouvelles possibilités pour étudier la dynamique complexe de cette famille de protéines. L'objectif de notre « grand défi » était d'établir un nouveau protocole capable de caractériser le paysage énergétique du récepteur de la ghréline, un récepteur typique de la classe A des RCPGs et de comprendre sa modulation par différents ligands.

600 000 heures GPU.

Les récepteurs couplés aux protéines G (RCPGs) constituent une grande famille de protéines membranaires de plus de 800 membres essentiels à la communication cellulaire. Présents à la surface de nombreuses cellules humaines, les RCPGs réagissent à un large éventail de signaux provenant de l'extérieur de la cellule, notamment des hormones, des neurotransmetteurs et d'autres stimuli sensoriels tels que la lumière, le goût et l'odorat.
Lorsqu'une molécule spécifique (ligand) se lie à un RCPG, elle module l'activation du récepteur en lui faisant subir d'importants changements de conformation. Ce changement de conformation permet au récepteur d'interagir avec d'autres protéines à l'intérieur de la cellule, déclenchant ainsi une série d'événements conduisant à des réponses bien spécifiques.
Cette cascade de signalisation est vitale pour de nombreux processus physiologiques, par exemple la régulation du rythme cardiaque, de la pression artérielle, etc... Ils jouent également un rôle clé dans le système nerveux central, affectant l'humeur, la perception et le comportement.

Pour ces raisons, les RCPGs sont la cible d'environ 30 % des médicaments actuellement sur le marché, combattant notamment l'hypertension, l'asthme ou encore la maladie de Parkinson. Cependant, malgré le succès des thérapies actuelles ciblant les RCPGs, de nombreuses maladies impliquant un de ces récepteurs manquent encore de traitements efficaces, comme par exemple certains types de cancer, de troubles neurologiques ou de maladies métaboliques, mettant ainsi en évidence un besoin permanent de développer des médicaments innovants.

ÉTABLIR UN NOUVEAU PROTOCOLE CAPABLE DE CARACTÉRISER CES PAYSAGES ET LEUR MODULATION

La conception de médicaments assistée par ordinateur, en particulier dans le contexte des progrès récents des algorithmes basés sur l'IA, peut accélérer de manière significative l'identification de médicaments candidats potentiels interagissant avec les RCPGs. Toutefois, à ce jour, ces algorithmes reposent encore largement sur des structures expérimentales ou sur des modèles prédits par *AlphaFold*.

À SAVOIR
« Cette vision dynamique est la clé pour améliorer notre compréhension du processus biologique fondamental régissant la signalisation des RCPGs.»

> *« Le monde de l'infiniment petit se transforme en un monde de calculs. »*
>
> James GLEICK

Nous prévoyons d'étendre l'utilisation de notre protocole pour capturer l'effet de ces principaux partenaires protéiques intracellulaires.

Bien que ces approches fournissent le niveau requis d'informations structurelles, elles ne donnent qu'une vision statique de ces récepteurs. Or, il est largement admis que la signalisation des RCPGs est régie par un large spectre d'états conformationnels en équilibre dynamique. Dans ce contexte, l'action des ligands et des protéines partenaires des RCPGs est de moduler ces équilibres, donnant lieu à des paysages conformationnels complexes.

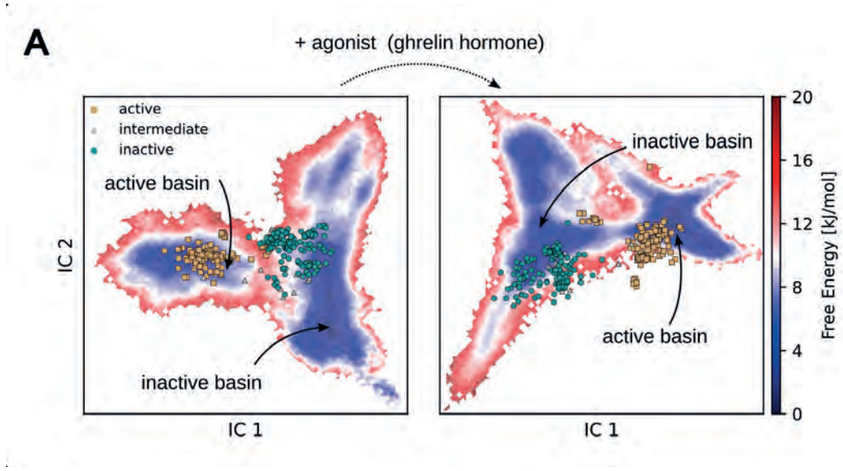
Durant la phase du Grand Challenge d'Adastra, notre projet a été d'établir un nouveau protocole capable de caractériser ces paysages et leur modulation par différents ligands et/ou protéines partenaires. Nous pensons que cette vision dynamique de ces systèmes est la clé pour améliorer notre compréhension du processus biologique fondamental régissant la signalisation des RCPGs et pourrait aider à concevoir des médicaments plus sélectifs ciblant ces protéines.
En utilisant des codes dédiés sur la machine Adastra, nous avons employé des simulations de dynamique moléculaire (DM) pour prédire la dynamique du récepteur de la ghréline (un RCPG typique de la classe A) sur des échelles de temps rarement atteintes auparavant.

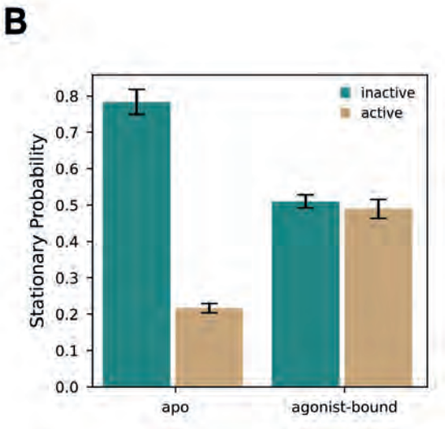
SIMULATIONS DE DYNAMIQUE MOLÉCULAIRE
Les simulations de dynamique moléculaire calculent les forces appliquées aux atomes pour prédire leurs mouvements au cours du temps en résolvant les équations de Newton.

MÉTA-DYNAMIQUE
La méta-dynamique est une technique de simulation biaisée où les mouvements du système sont favorisés dans des directions bien définies.

CAPTURER LES PRINCIPAUX ÉTATS CONFORMATIONNELS DU RÉCEPTEUR

En effet, la portabilité de codes comme *Gromacs* sur de telles machines à base de GPUs a considérablement boosté notre domaine de recherche ces dernières années. Nous avons testé puis validé notre protocole combinant des simulations de méta-dynamique par échange de biais, la DM libre classique et des analyses en chaînes de Markov pour explorer le paysage conformationnel du récepteur de la ghréline.
Nos simulations ont non seulement permis de capturer les principaux états conformationnels du récepteur, mais aussi leur cinétique de transition. Les simulations réalisées avec et sans l'hormone ghréline ont également permis de saisir comment ce peptide modifie le paysage énergétique de son récepteur, en abaissant ou en augmentant les énergies de certaines de ces conformations.





(A) Paysage d'énergie libre du récepteur apo (à gauche) et lié à la ghréline (à droite). Les structures expérimentales des RCPG de classe A sont présentées en fonction de leur état d'activation : actif (carrés jaunes), inactif (cercles verts) ou intermédiaire (triangles gris). (B) Les probabilités stationnaires des bassins d'énergie inactifs (verts) et actifs (jaunes) indiqués ont été calculées pour le système apo et lié à la ghréline.

L'EFFET D'AUTRES CLASSES DE LIGANDS SUR CE RÉCEPTEUR

En outre, nous avons identifié de nombreux sous-états méta-stables de ce récepteur non encore capturés par des structures expérimentales et qui pourraient s'avérer intéressants pour la conception de futurs médicaments.
En bon accord avec les données expérimentales de la littérature, les informations dynamiques que nous avons obtenues nous aideront certainement à mieux comprendre le mécanisme complexe d'activation des RCPGs.

Dans la continuité de ce travail, les calculs actuellement effectués sur Adastra se concentrent sur l'effet d'autres classes de ligands ciblant ce même récepteur.
En outre, nous prévoyons d'étendre l'utilisation de notre protocole pour capturer l'effet de ces principaux partenaires protéiques intracellulaires.
Cette prochaine étape nécessitera le test et la validation de modèles spécifiques dits « gros grains », plus adaptés à la taille accrue de ces systèmes et permettant une exploration beaucoup plus rapide de ces paysages conformationnels complexes. ■

CHAÎNES DE MARKOV
Les chaînes de Markov analysent les simulations atomiques pour prédire, à l'équilibre, les taux de transition entre les différents états du système.

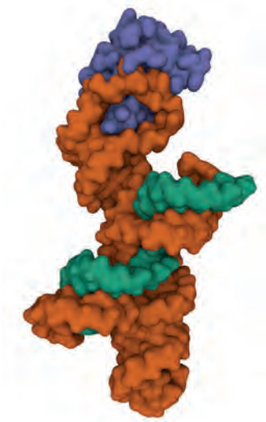
CT7 **MODÉLISATION MOLÉCULAIRE APPLIQUÉE À LA BIOLOGIE**

L'ÉQUIPE
Guillaume STIRNEMANN (ENS et CNRS, UMR8640 PASTEUR)
Damien LAAGE (ENS et CNRS, UMR8640 PASTEUR)
Sélène FORGET (ENS et CNRS, UMR8640 PASTEUR)
Paul GUENON (ENS et CNRS, UMR8640 PASTEUR)

x 100

C'est l'ordre de grandeur de l'accélération des simulations sur un cœur GPU et 10 cœurs CPU, comparé à un 10 cœurs CPU seuls.

Simulations haut-débit pour sonder les conformations biomoléculaires



LE CONTEXTE. La dynamique moléculaire permet de simuler l'évolution temporelle d'une macromolécule biologique. Très gourmande en ressources, elle ne permet pas de dépasser l'échelle de temps de la microseconde et de propager plus que quelques échantillons de façon routinière. La disponibilité de vastes ressources GPUs permet de faire sauter ce plafond de verre en accélérant l'échantillonnage et en permettant une réplication massive des trajectoires.

90 000 heures GPU.

Le fonctionnement du vivant découle des interactions entre les molécules qui constituent les cellules, en premier lieu desquelles les biomolécules, comme l'ADN ou les protéines. Alors que le premier sert à stocker l'information génétique, les secondes assurent la plupart des fonctions cellulaires.
Un dogme central en biologie est que la fonction d'une protéine découle directement de sa structure, c'est-à dire-de l'arrangement tri-dimensionnel des atomes qui constituent sa séquence. Il n'est donc pas étonnant qu'un des graals de la biologie moléculaire consiste à déterminer cette structure à l'aide de techniques expérimentales (diffraction des rayons X notamment).
Ceci explique aussi le retentissement médiatique récent (et la réelle révolution qu'ils apportent) qui accompagna la mise au point de techniques d'intelligence artificielle comme *AlphaFold2*, entrainées sur des structures existantes, pour prédire la structure de n'importe quelle protéine. Quoique cruciales, ces approches ne donnent en général qu'une image statique de la biomolécule, ne fournissant qu'une structure unique, ou au mieux une poignée de structures. Cela pose plusieurs problèmes.

Par exemple, il est maintenant bien établi que les états actifs d'une protéine peuvent correspondre à des états métastables, difficiles à isoler expérimentalement, ou par les méthodes d'IA. Dans d'autres cas, ces seules structures ne permettent pas de comprendre comment une perturbation du système peut déclencher un changement de forme et de fonction qui en découle (allostérie au sens large).

PREMIER TEMPS, ÉTUDIER L'EFFET ALLOSTÉRIQUE D'UNE MUTATION PONCTUELLE

Dans ce contexte, l'apport d'approches issues des simulations moléculaires peut s'avérer crucial. Mais, d'une part, ces simulations sont gourmandes en ressources, et d'autre part, les phénomènes moléculaires sont stochastiques : des dizaines d'observations sont nécessaires pour obtenir une image correcte d'un mécanisme et en déterminer la thermodynamique.
Grâce aux Grands Challenges, nous avons pu mener une approche de simulations à grande échelle dans le cadre de deux projets sur des systèmes distincts mais à la philosophie commune.

DYNAMIQUE MOLÉCULAIRE
Intégration numérique des équations du mouvement par pas de temps infinitésimaux (de l'ordre de la femtoseconde).

> *« Le tout est bien plus puissant que la somme des parties [et conduit à] une formidable accélération de l'exploration des conformations. »*

Cela ouvre la voie à un échantillonnage statistiquement efficace des systèmes biomoléculaires encore difficilement imaginable il y a quelques années.

Dans un premier temps, Paul Guénon, doctorant au laboratoire, a pu étudier l'effet allostérique d'une mutation ponctuelle sur une enzyme modèle. En pouvant générer cent trajectoires d'une microseconde chacune, qui auraient demandé plusieurs années de calculs sur une robuste machine de bureau, il a pu bénéficier de l'utilisation simultanée de dizaines de nœuds GPU pour effectuer ces calculs en quelques jours. Ses simulations ont permis de révéler le mécanisme moléculaire de l'allostérie latente de ce système, c'est-à-dire l'enchaînement des événements microscopiques qui conduisent à un changement structurel majeur de la protéine après mutation.

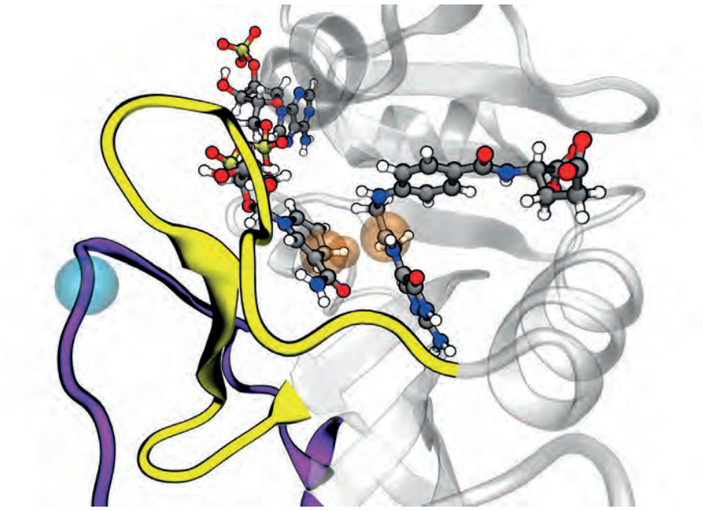
ALLOSTÉRIE
Régulation de l'activité protéique en un site par une perturbation sur un autre site distant.

En analysant les résultats de ses simulations dans le cadre d'un modèle de la physique statistique appelé «à chaînes de Markov», il a pu extraire les états structurels caractéristiques rencontrés le long de la transition et les constantes de vitesse correspondant à chacune des étapes de la transition. Ces résultats fournissent un éclairage nouveau sur les mécanismes de communication à longue portée entre des sites distants au sein d'une protéine, au cœur des phénomènes allostériques.

RIBOZYME
ARN capable de catalyser (c'est à dire accélérer) des réactions chimiques, souvent sa propre formation ou son clivage.

UN RÔLE CATALYTIQUE DE CERTAINS ARNs

Une autre doctorante du laboratoire, Sélène Forget, étudie, elle, l'ARN. Connu des biologistes depuis longtemps, l'ARN fut récemment mis en lumière par l'utilisation de vaccins révolutionnaires la ciblant, qui ont en grande partie permis d'enrayer l'épidémie de Covid19.
Alors qu'on a longtemps cru que le rôle de l'ARN se cantonnait à la traduction du code génétique contenu dans l'ADN sous forme de protéines (c'est d'ailleurs de tels ARNs messagers qui sont ciblés par ces vaccins), une réalité bien plus complexe est progressivement mise au jour depuis les années 80, avec des rôles très variés et cruciaux pour le vivant, notamment dans l'expression génétique.
Mais une des découvertes les plus spectaculaires concerne un rôle catalytique de certains ARNs, au même titre que les enzymes de protéines qui sont beaucoup plus répandues. Ces ribozymes ont un rôle physiologique chez certains eucaryotes, mais surtout dans certains virus ; et la machinerie à l'origine de la synthèse des protéines, le ribosome, est essentiellement constituée d'ARN dans son centre réactif.



Effet allostérique de la mutation d'un résidu (boule bleue) sur la réactivité d'une enzyme.

UNE TECHNIQUE DE SIMULATION VISANT À AMÉLIORER L'EXPLORATION DE L'ESPACE CONFORMATIONNEL

De façon cruciale, on pense aussi désormais que ces ribozymes auraient pu être à l'origine des premiers systèmes vivants primitifs, probablement basés essentiellement sur de l'ARN. Cela démontre l'intérêt majeur qu'il y a à comprendre comment fonctionnent ces petits réacteurs biomoléculaires.
Les structures expérimentales disponibles ne fournissent pas forcément une image fiable du paysage de structures biologiquement actives, notamment car l'ARN est beaucoup plus flexible et malléable que les protéines : une image statique ne saurait en dépeindre toute la complexité.

Dans ce cadre, nous avons tiré profit de l'utilisation de dizaines de GPUs en parallèle pour mettre en place une technique de simulation visant à améliorer l'exploration de l'espace conformationnel, mais très gourmande en ressources.
Cette technique, les simulations répliquées à échange d'Hamiltonien, consiste en la propagation simultanée de dizaines de répliques du système, chacune simulée dans des conditions thermodynamiques légèrement différentes. Grâce à la magie des lois de la physique statistique, le tout est bien plus puissant que la somme des parties ; c'est-à-dire que la communication entre répliques permet une formidable accélération de l'exploration des conformations.

DÉCOUVRIR ET CARACTÉRISER DE NOUVELLES CONFORMATIONS DU RIBOZYME

Autrement dit, l'aspect crucial ici est de disposer de ressources conséquentes travaillant en parallèle, plutôt qu'une seule ressource travaillant de façon continue pendant plus longtemps. Ces simulations ont permis de découvrir et de caractériser de nouvelles conformations du ribozyme étudié, ce qui impacte directement la compréhension de sa réactivité.

En conclusion, la disponibilité de ressources GPU en quantité, bien adaptés aux calculs nécessaires à la simulation moléculaire, et en grande quantité, permet une propagation massive de trajectoires et ouvre la voie à un échantillonnage statistiquement efficace des systèmes biomoléculaires encore difficilement imaginable il y a quelques années sur des ressources plus classiques et/ou limitées. ■

CT10 **APPLICATIONS TRANVERSES & NOUVELLES APPLICATIONS**

L'ÉQUIPE
Florent BARTOCCIONI (Valeo.ai, Centre Inria-Univ. Grenoble Alpes)
Yihong XU (Vaeo.ai)
Mickaël CHEN (Vaeo.ai)
Eloi ZABLOCKI (Vaeo.ai)

Tuan-Hung VU (Vaeo.ai)
Patrick PEREZ (Vaeo.ai)
Karteek ALAHARI (Centre Inria-Univ. Grenoble Alpes)

20 mètres

Les performances en prédiction sont optimales pour les agents à moins de 20m. Au-delà, elles se dégradent.

Détecter les usagers de la route, prédire leurs trajectoires

LE CONTEXTE. Dans la conduite autonome, détecter les agents et prévoir leurs mouvements est crucial pour la sécurité. Jusqu'à présent, ces deux tâches ont été développées de manière indépendante l'une de l'autre. Leur intégration dans un système unique, en vue d'un déploiement réel, laisse apparaître des défis : comment interagissent ensemble ces méthodes développées séparément ? Comment optimiser des ressources limitées pour améliorer la prédiction ?

240 000 heures.

Dans une chaîne de traitement dite conventionnelle, les données issues des capteurs (caméras, LiDARs, …) sont d'abord traitées par un module de perception chargé de détecter et localiser dans l'espace les agents environnants. Ceux-ci sont représentés dans un espace commun dit «en vue d'oiseau». Dans un second temps, les localisations ainsi estimées sont utilisées par un module de prédiction de trajectoire pour tenter d'anticiper leurs déplacements potentiels. Toutes ces informations sont finalement rendues exploitables pour un modèle de planification qui va proposer les actions futures de l'égo-véhicule. Jusque récemment, la littérature scientifique s'est attaquée à ces différentes tâches en silo, en élaborant indépendamment des standards sur chacune des sous-tâches. Ce découpage a permis des progrès rapides, en particulier en termes de performances prédictives individuelles des modèles.
Cependant, afin d'aller vers un déploiement dans un véhicule, cette performance ne suffit pas et il faut répondre à d'autres critères. Dans ce challenge, nous proposons une importante campagne d'évaluation des modèles de détection et de prédiction de trajectoires dans l'optique d'apporter de nouveaux éclairages et se rapprocher d'un déploiement.

UN NOUVEAU PARADIGME DE PRÉDICTION DE TRAJECTOIRE...

Tout d'abord, la capacité embarquée dans un véhicule étant limitée, il faut prendre en considération le coût computationnel nécessaire à une utilisation en temps réel. Ainsi, le premier volet de notre étude s'intéresse à l'évaluation des rapports coût computationnel / bénéfice en performance prédictive des modèles de détections. On s'intéresse par exemple à l'impact de la résolution des caméras et de la taille de différents blocs dans le modèle. Nous établissons ainsi les rapports entre le coût en mémoire, la vitesse d'inférence, et la précision des prédictions. Ces données permettront d'éclairer sur les choix de déploiement en tenant compte des besoins en performance, du coût de la capacité de calcul embarqué, ainsi que des performances des capteurs.

Ensuite, la question de l'intégration des modules dans une chaîne de traitement complète n'est que rarement abordée en recherche. Or, le modèle de prédiction de trajectoire est conçu, entraîné et évalué à partir de trajectoires annotées : le modèle a

DONNÉES CAPTEURS
Les véhicules peuvent être équipés de caméras et/ou de LiDARs. Les modèles fusionnent la donnée pour extraire l'information.

VUE D'OISEAU
Un repère en vue du dessus, centrée autour de l'égo-véhicule. Les modèles y projettent et traitent les données disponibles.

> *Il est crucial de mesurer l'adéquation entre les pratiques en recherche et les contraintes du monde réel.*

Parmi nos conclusions, nous mettons en évidence la très faible performance des modèles de prédiction de trajectoires une fois intégré à la chaîne de traitement.

Exemple de détection.
À partir des données capteurs (à gauche), le système de perception produit une carte des éléments environnants (à droite). `GT' indique la visualisation de la vérité terrain.

BOUT-EN-BOUT
Paradigme où le système traite directement les données brutes des capteurs jusqu'à la prédiction finale. Le module de prédiction de trajectoire a donc des entrées réelles et bruitées (erreur de détection et/ou de positionnement).

alors accès à des données de bonne qualité et peut s'appuyer sur leur fiabilité.
À l'inverse, dans des conditions de déploiement, les trajectoires observées sont prédites par un modèle imparfait et biaisé, plus ou moins fiable. Le comportement des modèles en déploiement peuvent donc être très différents de leur comportement sous les conditions de conception et d'entraînement.

Palliant à cette disparité, un nouveau paradigme de prédiction de trajectoire dit de bout-en-bout, commence à émerger, mais les différences méthodologiques complexifient la comparaison avec le paradigme conventionnel.

UNE MÉTHODOLOGIE D'ÉVALUATION JOINTE DES MODULES DE PERCEPTION ET DE PRÉDICTION

Dans le détail, la principale différence qui pose problème est qu'en prédiction de trajectoire *conventionnelle*, on cherche à prédire le futur pour un agent donné alors que dans le paradigme de bout-en-bout, les agents ne sont pas connus à l'avance mais sont plutôt une sortie du système. Certains peuvent ne pas avoir été détectés tandis que d'autres peuvent être hallucinés. En conséquence, il n'est pas possible d'évaluer les modèles de bout-en-bout de manière conventionnelle.
En mettant au point une méthodologie d'évaluation jointe des modules de perception et de prédiction de trajectoire, nous proposons une analyse comparée de ces deux paradigmes. De plus, nous caractérisons l'impact sur la tâche de prédiction de trajectoire de différents types de problèmes rencontrés en détection.

PROPOSER DES MÉTHODOLOGIES ET DES MÉTRIQUES D'ÉVALUATIONS PLUS EN PHASE

Parmi nos conclusions, nous mettons en évidence la très faible performance des modèles de prédiction de trajectoires une fois intégré à la chaîne de traitement, soulignant l'importance de prendre ces aspects en considération. De façon plus étonnante, nous soulignons aussi que jusqu'à présent, ni les stratégies simples d'adaptation, ni le paradigme de *bout-en-bout*, ne semblent efficaces face à ce problème.
Notre étude détaillée caractérise, par ailleurs et entre autres l'impact de l'utilisation de capteur *LiDAR* plutôt que d'une caméra, de la présence de faux positifs ou faux négatifs du module de détection et de l'importance de la précision de la localisation des agents.
Pour les modules de détection, ces données nous permettent de proposer des méthodologies et des métriques d'évaluations plus en phase avec les besoins du module de prédiction de trajectoire.
Nous soumettons aussi les points d'amélioration possible des méthodologies employées en prédiction de trajectoire, tel que la prise en compte des limites réalistes en termes de qualités et performances des modules de détection pour des objets situés au-delà de 20 mètres.

Nos résultats sur le premier volet ont été ajoutés à l'appendice de [1]. Le deuxième volet de l'étude est à l'origine du projet ayant mené à une publication [2]. Les codes source sont aussi disponibles publiquement. ■

[1] LaRa: Latents and Rays for Multi-Camera Bird's-Eye-View Semantic Segmentation
Florent Bartoccioni, Éloi Zablocki, Andrei Bursuc, Patrick Pérez, Matthieu Cord, Karteek Alahari, CoRL 2022.
[2] Towards Motion Forecasting with Real-World Perception Inputs: Are End-to-End Approaches Competitive?
Yihong Xu, Loïck Chambon, Éloi Zablocki, Mickaël Chen, Alexandre Alahi, Matthieu Cord, Patrick Pérez, ICRA 2024.

AGENTS
Les différents participants du trafic (véhicules, piétons...) dont le système doit prédire les trajectoires par rapport à l'égo-véhicule.

LIDAR
Systèmes *(Ligth detection and ranging)* qui émettent une lumière laser à partir de divers systèmes mobiles (automobiles, avions, drones...) à travers l'air et la végétation (laser aérien) et même l'eau (laser bathymétrique). Un scanner reçoit la lumière en retour (échos) et mesure les distances et les angles.

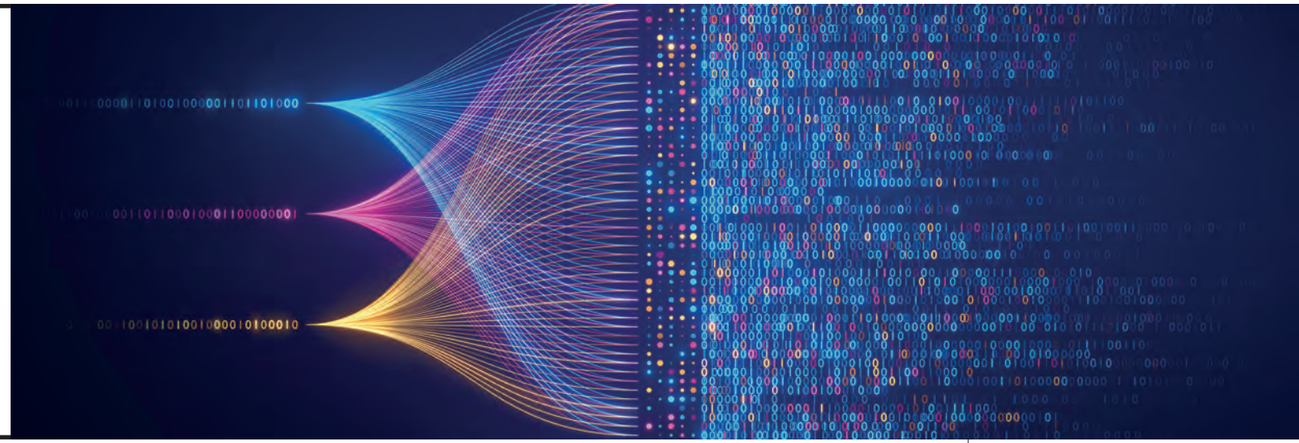
CT10 **APPLICATIONS TRANVERSES & NOUVELLES APPLICATIONS**

L'ÉQUIPE
Matthieu CORD (Sorbonne Univ.)
Mustafa SHUKOR (Sorbonne Univ.)
Corentin DANCETTE (Sorbonne Univ.)

250 millions

De paramètres. Notre modèle obtient des performances compétitives par rapport aux meilleures approches dédiées à des modalités spécifiques, mais aussi par rapport à des modèles génériques jusqu'à 100 fois plus grands.

Image, langage, son : un seul modèle pour tous !



LE CONTEXTE. Il est aujourd'hui essentiel de développer des approches multimodales vision-audio-langage capables de résoudre une variété de tâches telles que la recherche image-texte, le légendage d'images et la réponse à des questions visuelles (VQA) . La stratégie principale consiste à construire un modèle unifié qui peut traiter toutes les tâches ensemble.

200 000 heures.

Grâce aux grands modèles de langage (LLMs) , la quête d'agents généralistes est loin d'être un fantasme. L'un des principaux obstacles à la construction de ces modèles généraux est la diversité et l'hétérogénéité des tâches et des modalités. Néanmoins, leur limitation actuelle à une seule modalité (le texte) restreint leur compréhension et leur interaction avec le monde réel. Malgré cette limitation, les LLMs sont devenus un composant essentiel dans de nombreux systèmes, en raison du transfert de la représentation du langage vers d'autres modalités. Récemment, des travaux ont tenté d'aller au-delà de la modalité unique et de construire des modèles multimodaux puissants qui surpassent les approches spécifiques à une tâche ou à une modalité. Cependant, la plupart de ces travaux se concentrent sur des tâches image-texte et seule une poignée d'approches vise à intégrer plus de deux modalités, telles que l'image/vidéo/texte. Quelques travaux ont été réalisés dans le domaine des tâches vidéo/texte, et l'exploration du domaine des tâches audio-texte reste très limitée. Seule une poignée d'approches ont cherché à intégrer plus de deux modalités, telles que image/vidéo/texte ou, rarement, image/vidéo/audio/texte.

Une solution prometteuse est l'unification, permettant la prise en charge d'une myriade de tâches et de modalités dans un cadre unifié. Cela souligne la nécessité de disposer de modèles multimodaux robustes capables de prendre en charge diverses tâches à travers de nombreuses modalités.

EXPLORER LE PASSAGE À L'ÉCHELLE DE DIFFÉRENTS MODÈLES MULTIMODAUX

L'objectif de notre travail vise à développer des approches multimodales efficaces qui peuvent résoudre une variété de tâches telles que la recherche Image-texte, la génération de légendes d'image et la tache de question/réponse visuelle, ainsi que des extensions audio et vidéo.

Nous nous concentrons sur la construction d'un modèle unifié pouvant gérer efficacement toutes les tâches. Notre modèle *UnIVAL* illustré sur la Figure 1 offre un cadre généraliste et agnostique aux types de modalités. Il va au-delà de deux modalités, unifiant texte, images, vidéo et audio en un seul modèle.

Le succès de ces grands modèles est principalement dû à leur pré-entraînement sur de très grandes bases de données. Cela est donc très coûteux en ressource (annotation et/ou préparation des données, et calcul GPU), limitant drastiquement

DONNÉES CAPTEURS VQA (VISUAL QUESTION ANSWERING)
C'est une tâche d'intelligence artificielle où un modèle doit répondre à des questions en langage naturel en se basant sur le contenu visuel d'une image. Cette tâche combine la compréhension d'images et l'analyse linguistique pour fournir des réponses pertinentes.

GRANDS MODÈLES DE LANGAGE (LLMS)
Ce sont des systèmes d'intelligence artificielle entraînés sur d'énormes ensembles de données textuelles pour comprendre, générer et manipuler le langage. Ils sont utilisés pour une variété de tâches, telles que la traduction, la génération de texte et la réponse aux questions.

« Un modèle unique pour traiter tous les types de données nécessite de les plonger dans un espace de travail commun difficile à définir mathématiquement. »

Nous démontrons les avantages de l'apprentissage avec curriculum multimodal en équilibrant les tâches pour entraîner efficacement le modèle au-delà de deux modalités.

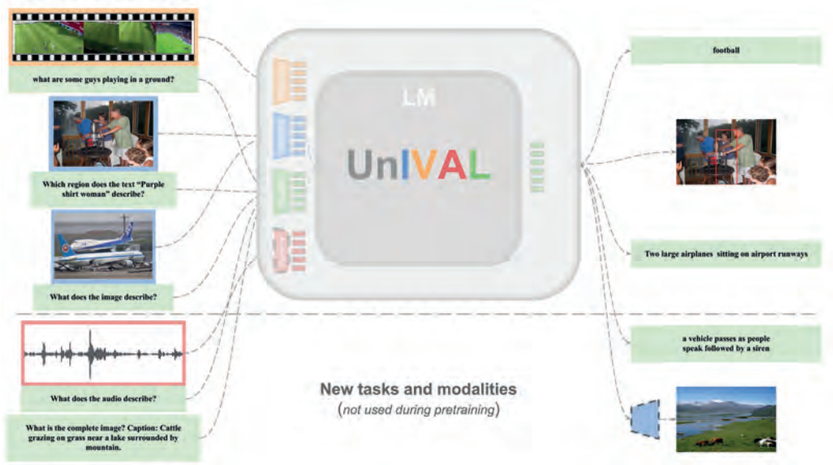


Figure 1 :
UnIVAL est une architecture unifiée avec plusieurs tâches et différents formats d'entrée/sortie. Notre modèle UnIVAL est entraîné sur des tâches d'image et de vidéo-texte et peut traiter de nouvelles modalités (audio-texte) et tâches (génération texte-image) qui n'ont pas été vues lors de l'entraînement.

MULTIMODAL
Type d'apprentissage qui intègre plusieurs types de données appelées modalités, telles que le texte, l'audio, les images ou la vidéo.

les possibilités d'investigation de ces approches pour les laboratoires académiques. Notre ambition dans ce projet est d'explorer le passage à l'échelle de différents modèles multimodaux afin de fournir à la communauté scientifique une meilleure vision sur :
1/ les architectures intéressantes, **2/** les stratégies d'apprentissage, **3/**l'efficacité et les manières d'apprendre plus économes. Les moyens de calcul exceptionnels mis à notre disposition dans le cadre de ce Grand Challenge nous ont permis de réaliser tous les apprentissages de nos modèles multimodaux, ainsi que les différents tests et comparaisons nécessaires.

DEUX PREMIÈRES CONTRIBUTIONS VALIDÉES...

Les contributions que nous avons validées grâce aux nombreuses expériences sont multiples.
> *UnIVAL* est le premier modèle avec une architecture, un format d'entrée/sortie et un objectif d'entraînement unifiés, capable de traiter des tâches de langage, d'images, de vidéos et d'audio sans recourir à un entraînement à trop grande échelle.
> Nous démontrons les avantages de l'apprentissage avec curriculum multimodal en équilibrant les tâches pour entraîner efficacement le modèle au-delà de deux modalités. Plus précisément, nous avons mis au point deux stratégies pour améliorer l'efficacité du modèle *UnIVAL* lors de l'apprentissage. La première, *--Multimodal Curriculum Learning (MCL)* vise à introduire progressivement des modalités supplémentaires au cours de l'entraînement, facilitant une transition en douceur et réduisant les besoins en calcul. La seconde, *--Multimodal Task Balancing* permet d'équilibrer les tâches au sein des batchs d'entraînement, optimisant les performances globales, notamment en présence d'ensembles de données déséquilibrés. Ces méthodes visent à améliorer à la fois l'efficacité computationnelle et les performances du modèle.

... SUIVIES DE DEUX AUTRES CONTRIBUTIONS VALIDÉES

Parmi nos autres contributions :
> Nous validons l'importance de l'entraînement multitâche par rapport à l'approche standard et étudions le transfert de connaissances entre les tâches et les modalités pré-entraînées. L'entraînement sur un plus grand nombre de modalités améliore la capacité du modèle à généraliser à de nouvelles modalités. Par exemple, sans aucun pré-entraînement audio, *UnIVAL* peut atteindre des performances compétitives par rapport à l'état de l'art lorsqu'il est appliqué sur des tâches audio-texte.
> Nous proposons une nouvelle étude sur la fusion de modèles multimodaux via l'interpolation de poids des différents réseaux appris.
Nous montrons que lorsque les poids sont réglés sur différentes tâches multimodales à partir de notre modèle pré-entraîné unifié, l'interpolation dans l'espace des poids peut combiner efficacement les compétences des différents réseaux, créant des modèles multitâches plus robustes sans surcharge de calcul à l'inférence.
Il s'agit de la première étude d'interpolation de poids montrant son efficacité avec des grands modèles multimodaux.

UNE ÉTAPE SIGNIFICATIVE VERS DES SYSTÈMES GÉNÉRIQUES ET MULTIMODAUX

Notre modèle *UnIVAL* doit encore surmonter certains défis, notamment la gestion des données bruitées et la nécessité de capacités de calcul accrues dans certaines configurations. L'intégration fluide de nouvelles modalités et de nouvelles tâches reste également un domaine de recherche actif.
Plusieurs applications sont envisageables pour *UnIVAL*, comme l'amélioration des systèmes d'assistance virtuelle, la création de contenu multimodal enrichi et l'amélioration des outils de traduction automatique. L'évolution des modèles unifiés est traitée ici dans le contexte de la communauté de l'IA, en s'inspirant des avancées dans le traitement du langage naturel (NLP).
UnIVAL représente une étape significative vers des systèmes génériques et multimodaux, capables de s'adapter à un large éventail de tâches. Plusieurs axes de recherche futurs, notamment l'amélioration de l'efficacité de l'entraînement, l'extension des capacités de traitement multimodal et l'exploration de nouvelles architectures de modèles sont possibles.
Enfin, il faut souligner également l'importance de collaborations interdisciplinaires pour surmonter les défis techniques et éthiques associés à l'utilisation de tels modèles.L'ensemble de ce travail a été publié dans le journal TMLR et le code ainsi que les modèles sont disponibles à l'adresse suivante : **https://unival-model.github.io/.** ■

TMLR
Revue dans le domaine du machine learning
https://jmlr.org/tmlr/

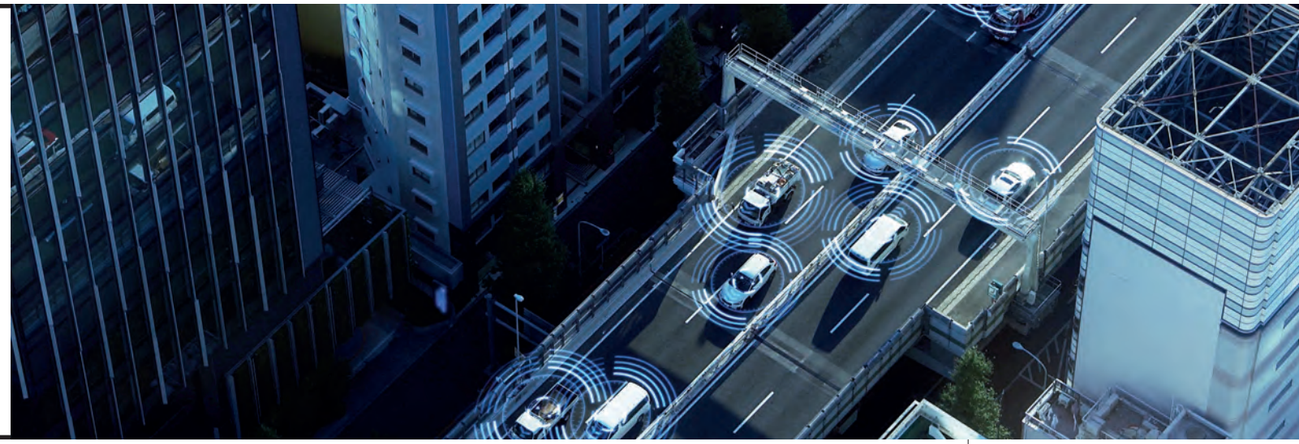
CT10 **APPLICATIONS TRANVERSES & NOUVELLES APPLICATIONS**

L'ÉQUIPE
Spyros GIDARIS (Valeo.ai)
Oriane SIMEONI (Valeo.ai)
Gilles PUY (Valeo.ai)
Alexandre BOULCH (Valeo.ai)

Corentin SAUTIER (Valeo.ai)
Andrei BURSUC (Valeo.ai)
Patrick PEREZ (Kyutai)
Renaud MARLET (Valeo.ai)

3 Notre méthode d'auto-supervision permet de diviser par 3 l'écart avec l'apprentissage supervisé.

Modèles de fondation *LiDAR* pour la conduite autonome



LE CONTEXTE. Les capteurs *LiDAR* capturent une représentation 3D du monde, complétant ainsi efficacement les informations obtenues par les caméras pour renforcer la sécurité des voitures autonomes. Toutefois, l'entraînement des modèles de réseaux de neurones profonds pour le *LiDAR* nécessite de vastes campagnes d'annotation qui nécessitent beaucoup de temps et d'argent. Pour pallier cela, nous exploitons les modèles de fondation 2D pour les images prêtes à l'emploi, afin de pré-entraîner les modèles *LiDAR* sur de grandes bases de données non annotées.

230 000 heures.

Le *LiDAR* est un capteur qui utilise des lasers pour créer une «photo» 3D de l'environnement qui l'entoure. Il fonctionne en mesurant la distance entre le capteur et les objets environnants, produisant ainsi un nuage de points 3D détaillé de la scène. La technologie *LiDAR* est essentielle pour l'amélioration des systèmes d'assistance à la conduite car elle permet d'obtenir des cartes 3D de l'environnement autour du véhicule, même dans des conditions de faible luminosité, par exemple durant la nuit.
Une manière d'exploiter les données produites par un capteur LiDAR est d'assigner une étiquette sémantique, tel que «véhicule», «obstacle», ou «piéton», à chaque point du nuage 3D. Ce processus, appelé segmentation sémantique de *LiDAR*, permet au véhicule autonome de localiser précisément, par exemple, les obstacles autour de lui.

Les meilleures méthodes pour la segmentation sémantique de nuages de points utilisent des réseaux de neurones profonds, également appelés modèles, qui sont entraînés sur une grande quantité de données annotées manuellement. Cependant, annoter les nuages de points issus d'un *LiDAR* est une tâche chronophage et très coûteuse.

DÉVELOPPER UNE TECHNIQUE D'APPRENTISSAGE AUTO-SUPERVISÉ SPÉCIFIQUE

Une approche moins gloutonne consiste à utiliser un pré-entraînement auto-supervisé. Cela consiste à entraîner un modèle avec une tâche prétexte en utilisant un ensemble de données non annotées. Le réseau ainsi pré-entraîné peut ensuite être affiné *(finetuned en anglais)* pour diverses tâches en aval.
Cette approche permet d'atteindre les mêmes performances que les modèles entièrement supervisés avec beaucoup moins d'annotations, voire de surpasser les performances de ces modèles avec la même quantité d'annotations. Il a été démontré que ce type de pré-entraînement auto-supervisé sur des images est très efficace pour apprendre des modèles génériques qui fonctionnent bien pour de nombreuses tâches en aval dans le domaine de l'image. Ces modèles sont appelés modèles de fondation 2D.

Dans ce projet, nous montrons que ces modèles de fondation 2D peuvent également être utilisés pour créer des modèles de fondation pour le *LiDAR*.
En effet, nous avons développé une

AFFINÉ
Désigne le processus de spécialisation d'un modèle général avec des données spécifiques, par exemple dans le domaine automobile, bancaire, énergétique, chimie...

MODÈLE DE FONDATION
Un grand réseau de neurones pré-entraîné sur des données non-annotées facilement adaptable à de nombreuses tâches.

« Passer à l'échelle, tant en taille de modèles que taille de base de données, est la clé pour améliorer les performances. »

Nous démontrons les avantages de l'apprentissage avec curriculum multimodal en équilibrant les tâches pour entraîner efficacement le modèle au-delà de deux modalités.

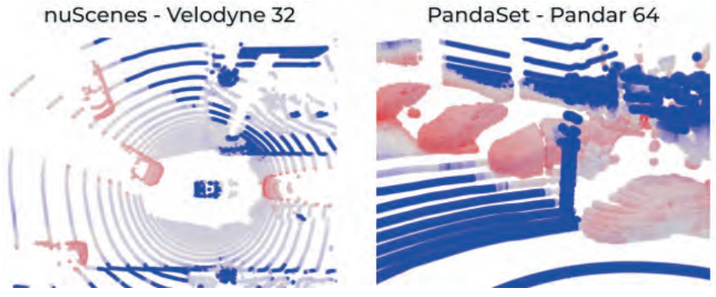


Figure 1 : Nos caractéristiques ScaLR, obtenues par auto-supervision, nous permettent de segmenter facilement les voitures (points rouge) dans divers types de données LiDAR issues de nuScenes [1] et PandaSet [2].

KNOWLEDGE DISTILLATION
La distillation des connaissances ou distillation de modèles est, dans le domaine de l'apprentissage automatique, le processus de transfert des connaissances d'un grand modèle, dit *« modèle enseignant »*, vers un modèle plus petit, dit *« modèle étudiant »*, plus facile à manipuler et à évaluer.

technique d'apprentissage auto-supervisée spécifiquement conçue pour la segmentation sémantique de nuages de points issus de *LiDAR* dans le contexte de la conduite autonome.
De nombreux véhicules de test sont équipés de caméras synchronisées et calibrées avec des capteurs *LiDAR*. Ces équipements fournissent une information panoramique de tout l'environnement autour du véhicule. Ces données sont bien plus faciles à acquérir qu'à annoter manuellement.

Notre approche consiste donc à exploiter ces données non annotées ainsi qu'un modèle de fondation 2D pour entraîner un réseau de neurones 3D dédié au *LiDAR*. Nous utilisons pour cela une technique d'apprentissage par transfert de connaissances *(knowledge distillation en anglais)*. Nous avons baptisé cette technique *ScaLR* [3].

AMÉLIORATION DES PERFORMANCES DES MODÈLES 3D AUTO-SUPERVISÉS

Comment cela fonctionne-t-il ? Dans une scène donnée, nous considérons une image et le nuage de points *LiDAR* acquis au même instant. Le réseau 3D prend en entrée le nuage de points et génère des caractéristiques (*features en anglais*) pour chacun des points 3D, tandis que le modèle de fondation 2D prend en entrée l'image et produit des caractéristiques pour chaque pixel de celle-ci.
En exploitant les correspondances géométriques entre les points 3D et les pixels de l'image, nous entraînons le réseau 3D à produire des caractéristiques identiques à celles du modèle de fondation 2D. Ainsi, le transfert de connaissance du modèle de fondation 2D vers le réseau 3D s'effectue sans nécessiter aucun effort humain.

Bien que des progrès récents aient été réalisés dans le transfert de connaissances entre les modèles 2D et 3D, un écart de performance significatif persiste entre les modèles 3D entraînés via cette technique et ceux entraînés entièrement sous supervision.

Dans le cadre de ce travail, nous avons amélioré les performances des modèles 3D auto-supervisés en examinant l'impact de trois principaux facteurs : la taille des modèles 3D, le modèle 2D pré-entraîné, et le jeu de données d'entraînement (non annoté).

DES MODÈLES PLUS ROBUSTES

Nous avons démontré que l'augmentation de la taille (le nombre de paramètres) des modèles 2D et 3D, ainsi que l'utilisation de données collectées à partir de différents *LiDAR* et dans diverses villes, permet d'améliorer considérablement la qualité des modèles 3D entraînés par transfert de connaissances.

L'utilisation de grands modèles 2D et 3D, nécessaire pour cette étude, a été rendue possible grâce aux ressources GPU allouées sur Adastra. Entre autres, sur la base de données pour la segmentation sémantique de nuages de points *LiDAR* appelée *nuScenes* [1], nous avons divisé par trois l'écart de performance entre les modèles 3D pré-entraînés par transfert de connaissances et ceux entraînés de manière entièrement supervisée.

De plus, nous avons également été capable de démontrer que les modèles entraînés avec notre méthode sont plus robustes que des modèles obtenus par supervision face aux changements environnementaux et aux perturbations au niveau du capteur *LiDAR*.

Dans l'esprit de la recherche collaborative en *Open source*, nous avons rendu notre code et les poids de nos modèles disponibles publiquement.

De cette manière, nous espérons contribuer au progrès collectif de notre domaine et permettre à nos collègues chercheurs du monde entier de s'appuyer sur notre travail (URL : **https://github.com/valeoai/ScaLR**). ■

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. CVPR, 2020.
[2] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, Yunlong Wang, and Diange Yang. PandaSet: Advanced sensor suite dataset for autonomous driving. ITSC, 2021.

APPRENTISSAGE AUTO SUPERVISÉ
(En anglais : *self-supervised learning, SSL*) Méthode d'apprentissage automatique où le modèle apprend à partir d'échantillons de données non annotées. Il peut être considéré comme une forme intermédiaire entre l'apprentissage supervisé et non supervisé.

RÉPERTOIRE DES PROJETS SCIENTIFIQUES

CT1
ENVIRONNEMENT

> Accélérer la simulation des tempêtes grâce aux GPU

CT4
GÉOPHYSIQUE & ASTROPHYSIQUE

> DynoStar, simulation à grande échelle de dynamo convective stellaire

> Simulations globales de disques d'accrétion à des températures extrêmes

> Imagerie sismique haute résolution en milieu viscoélastique

CT5
PHYSIQUE THÉORIQUE & PHYSIQUE DES PLASMAS

> Comprendre le rayonnement électromagnétique des sursauts radio solaires

CT7
MODÉLISATION MOLÉCULAIRE APPLIQUÉE À LA BIOLOGIE

> Irradiation ionisante de la matière biologique

> Exploration du paysage conformationnel du récepteur de la Ghréline

> Simulations haut-débit pour sonder les conformations biomoléculaires

CT10
APPLICATIONS TRANVERSES & NOUVELLES APPLICATIONS

> Détecter les usagers de la route, prédire leurs trajectoires

> Image, langage, son : un seul modèle pour tous !

> Modèles de fondation *LiDAR* pour la conduite autonome

Un grand merci aux contributeurs, aux rédacteurs et aux chercheurs pour le temps et l'énergie que vous avez consacrés à la réalisation de ce nouveau numéro des Grands Challenges.

Un grand merci également à l'ensemble de la communauté du calcul haute performance : les résultats de ces Grands Challenges soulignent aussi l'importance de son engagement dans la réussite de l'approche de GENCI.

Directeur de la publication : Philippe LAVOCAT - **Coordination :** Nicolas BELOT ; Annabel TRUONG - **Conception & réalisation :** avec des mots - **Couverture :** G. Cannat/CINES - **Photos & illustrations :** A. Delalande ; L. Bartoccioni ; DR ; N. Floquet/CINES ; iStock ; V. Monteiller ; F. Pantillon ; G. Stirnemann ; M. Van-Den-Bosshen - **Impression :** Quarante-Six - **GENCI** - 6, bis rue Auguste Vitu 75015 Paris / France - Tél. : +33 1 42 50 04 15

GENCI
6 bis, rue Auguste Vitu - 75015 Paris - France
Tél. : +33 1 42 50 04 15
www.genci.fr

Suivez GENCI