Focus

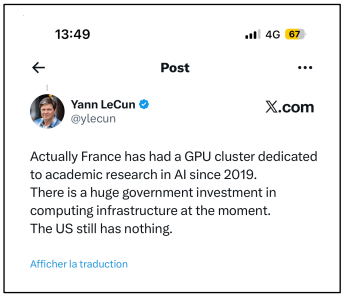
2025, année charnière pour l'Intelligence artificielle au service de la science

GENCI

Le calcul intensif au service de la connaissance

Tables des matières

Introduction



Le 9 février 2026, Yann LeCun énonçait sur son compte Twitter que la France dispose depuis 2019 d'un cluster GPU dédié à la recherche en Intelligence Artificielle (IA). Il soulignait l'effort important du gouvernement dans les infrastructures de recherche.

La pertinence du choix opéré par les pouvoirs publics nationaux à travers GENCI trouvait déjà une confirmation dans le rapport de la Cour des Comptes en 2023 : « *La mise en place d'un tel supercalculateur dans le cadre de la stratégie nationale en IA était une mesure pertinente, répondant à un réel besoin. Les résultats d'une consultation menée dans le cadre de l'enquête auprès de la communauté de chercheurs en IA confirment ce bien-fondé*. ».

Avec le supercalculateur Jean Zay mis à disposition depuis 2019 par GENCI à l'IDRIS, le centre de calcul du CNRS, les chercheurs, aussi bien dans le champ académique qu'industriel (startups, PME, grandes entreprises) ont accès gratuitement à ces ressources au titre de la recherche ouverte avec un accompagnement humain (support utilisateurs) mis en place par le CNRS dans le cadre du PNRIA (Plan National de Recherche en IA).

Après les annonces du Président de la République en 2023 à Vivatech, une quatrième extension de la machine Jean Zay a été acquise par GENCI auprès d'Eviden en 2024 comportant notamment 1 456 GPU H100 NVIDIA. Inaugurée en mai 2025, cette nouvelle partition confère à Jean Zay le statut de vaisseau amiral en matière d'IA au service de la recherche. En effet, Jean Zay est devenue, avec près de 1700 projets de recherche soutenus en 2025, la machine académique la plus utilisée en Europe pour la recherche recourant à l'IA. Ses ressources sont complétées depuis 2023 par les capacités déployées sur Adastra, au sein du CINES, avec plus de 1 500 GPU AMD MI250x et MI300A.

De très nombreux domaines de recherche et d'innovation bénéficient cette puissance accrue : recherche biomédicale, analyse de données astronomiques, mise au point de modèles pour la conduite autonome, conception de nouveaux matériaux, nouvelles énergies, agriculture, aide à la décision, culture, etc.

À gauche : le supercalculateur Adastra équipé d'accélérateurs MI250X. À droite, Adastra 2, équipe d'accélérateurs MI300A

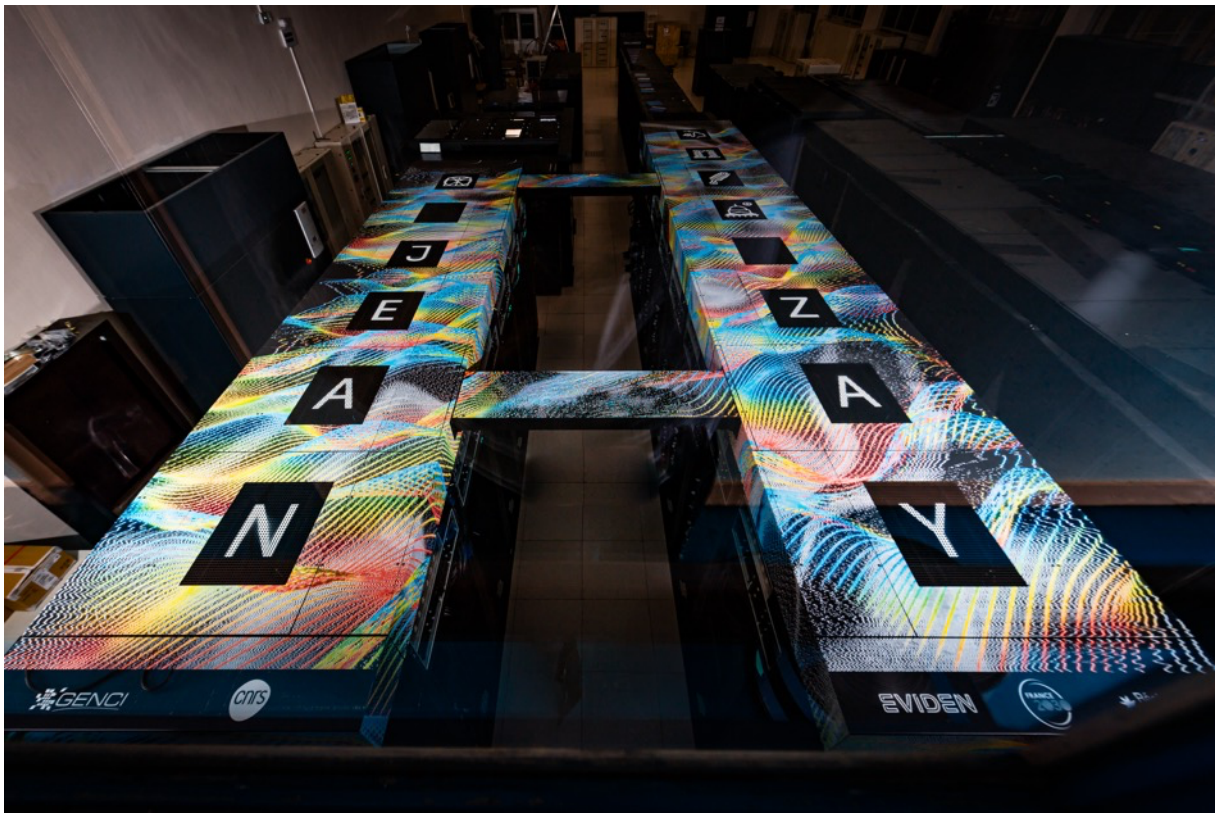
Plus spécifiquement concernant les projets en IA soutenus par GENCI sur ses moyens de calcul, il nous faut distinguer entre les ressources pour l'IA en tant qu'objet de recherche porté par le Comité Thématique 10, et celles consistant à mettre l'IA au service de l'activité scientifique portée maintenant par tous les autres Comités Thématiques (CT) scientifiques.

Ainsi 2025 a été la première année au cours de laquelle GENCI a alloué plus d'heures de calcul dans cette thématique dite « IA pour Science » que dans la thématique IA « classique » (incluant le traitement des langues, la vision, la robotique, l'IA agentique…). Ce résultat valide le choix fait en 2019 par GENCI et le CNRS de la convergence de ces deux typologies d'usages sur une même machine, catalysant le rapprochement entre la communauté de la simulation numérique et celle de l'intelligence artificielle.

2025 a également vu la mise sur pied du projet européen AI Factory France piloté par GENCI et Inria comme Agence de programme numérique, constituant une suite et une amplification majeure du dispositif mis en place. Elle formalise une ambition française et européenne pour la diffusion de l'IA au service de la science, de l'innovation et des services publics, notamment au service de 12 verticales métiers comme la santé, l'aéronautique, les sciences de la terre, les matériaux du futur, la défense, l'agriculture ou la robotique pour n'en citer que quelques-unes.

En fédérant des acteurs issus de la recherche et de l'innovation au sein des 19 *AI Factories* en Europe, AI Factory France offrira un point d'entrée unique (*one-stop shop*) pour la fourniture de d'infrastructures, de gestion des talents et de formation, d'expertise en HPC, IA et *datascience*, de mise à disposition de modèles, jeux de données et outils et enfin de services innovants pour les *startups* pleinement intégrés dans une vision européenne.

Le bilan de 2025 sur l'usage de ces ressources nationales en IA démontre que le pari de la convergence des technologies et des usages constitue une piste gagnante, piste qu'il importera de poursuivre en raison d'une demande désormais chroniquement supérieure à l'offre de calcul disponible. L'effet de saturation sur Jean Zay ayant en effet atteint un facteur cinq sur l'exercice écoulé.



Vue de la partie supérieure de la partition Jean Zay 4. Les éléments graphiques ont été conçus par le collectif Obvious

I De la convergence à la croissance

En 2025, GENCI a soutenu près de 1 700 projets en IA, avec une augmentation constante des heures attribuées et consommées. Les thématiques principales incluent la santé, l'énergie, et l'environnement. Les utilisateurs, industriels, startups et académiques, enthousiastes, anticipent une nouvelle augmentation de leurs besoins en ressources GPU.

- 1 693 projets ont eu recours aux ressources de GENCI en Intelligence Artificielle sur les supercalculateurs sur Jean Zay et Adastra.

- Sur l'ensemble des GPU mis à disposition par GENCI, la répartition entre académiques et industriels est la suivante :

- Industriels : 12 % des projets pour 21 % des heures
- Académiques : 88 % des projets pour 79 % des heures

- Mais, sur Jean Zay 4, la partition amirale, la proportion s'inverse passant de 12 à 70 % des heures de calcul consommées par des industriels, soit autour de 630 milles heures par an. Ces projets sont portés notamment par des entreprises comme Valeo, Linagora, Pyannote ou Huggingface, qui ont bénéficié de grands volumes d'heures de calcul.

Focus : à l'origine du supercalculateur Jean Zay

Le rapport sur l'intelligence artificielle (IA) rédigé par le mathématicien et député Cédric Villani, a été rendu public mercredi 28 mars 2018. Parmi les nombreuses pistes proposées, figuraient notamment : créer un réseau d'Instituts interdisciplinaires d'intelligence artificielle, mettre en place un supercalculateur conçu spécifiquement pour les applications d'IA, ou encore rendre plus attractives les carrières dans la recherche publique afin d'éviter la fuite des cerveaux vers les géants américains.

Dans le prolongement de l'annonce faite la même semaine par le Président de la République quant à la stratégie nationale en intelligence artificielle, le MESR (Ministère de l'Enseignement Supérieur et de la Recherche) a demandé à GENCI, opérateur en charge de mettre en œuvre la politique nationale d'équipement et d'accès à des moyens de calcul au service de la recherche, de répondre à l'une de ces recommandations: donner accès aux chercheurs français en intelligence artificielle (IA) à des moyens de calcul performants.

Deux principaux facteurs ont favorisé la mise en œuvre de cette opportunité :

- des échanges avec des représentants de la communauté de recherche française en IA ont permis de montrer que les besoins matériels de cette communauté pouvaient converger avec ceux des utilisateurs des ressources en simulation numérique (HPC). Cette communauté étant historiquement utilisatrice des moyens de calcul de GENCI depuis sa création en 2007.

- en 2018, les capacités de calcul de l'IDRIS (CNRS), l'un des 3 centres nationaux de calcul, étaient en cours d'évolution. Les besoins ainsi définis ont été intégrés dans l'appel d'offres que GENCI menait, dans le cadre de son plan pluriannuel d'investissements dans les 3 centres nationaux, pour équiper l'IDRIS courant 2019 avec un nouveau supercalculateur.

Cette machine inaugurerait un concept alors peu répandu en Europe : celui d'un supercalculateur *convergé,* c'est-à-dire répondant à la fois aux besoins des chercheurs en simulation numérique et à ceux en intelligence artificielle.

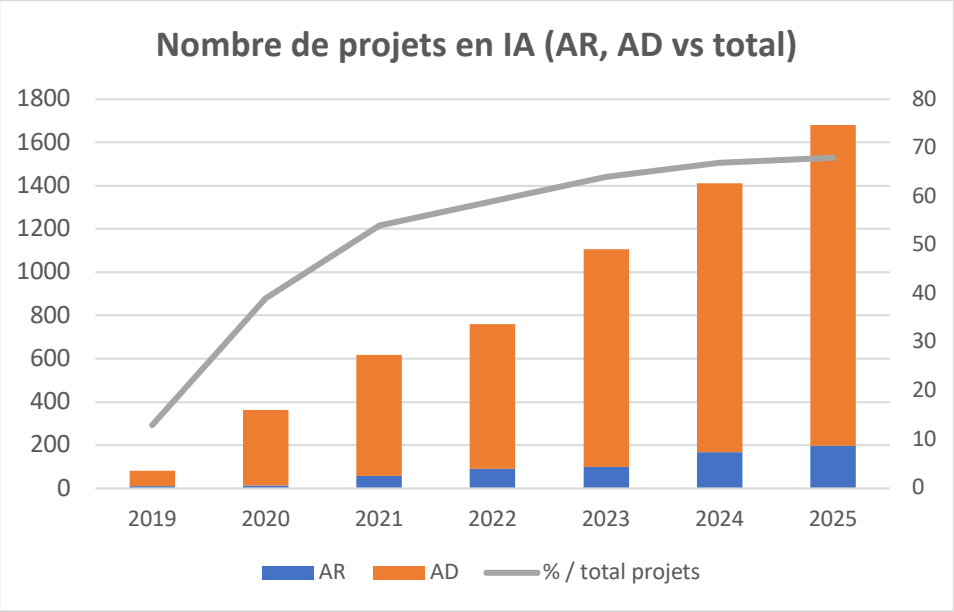
Outre l'optimisation technique et financière, ce dispositif technologique portait également une optimisation énergétique importante pour ce type de machines.

Plus encore, il contribuerait à rapprocher ces deux communautés. Au-delà de l'aspect technique, ce supercalculateur traduit aussi un projet mobilisant des ressources humaines donnant lieu à de nombreuses synergies.

Il a été nommé Jean Zay, en hommage à cette figure historique, ministre de l'Éducation nationale et des Beaux-Arts et cofondateur du CNRS

Portrait de Jean Zay © Studios Harcourt

L'IA comme objet de science et l'IA comme outil pour la science : Science for AI and AI for Science



Le graphique ci-dessus montre l'évolution des demandes pour des projets en IA depuis la mise en place de Jean Zay en 2019 selon les deux types d'accès mis à disposition par GENCI : accès dynamiques (AD) et accès réguliers (AR). En gris, figure le pourcentage de projets IA par rapport à tous les projets soutenus par GENCI.

- **Les accès dynamiques** sont ouverts toute l'année. Ils permettent en quelques clics et quelques jours d'avoir accès aux moyens de calcul de GENCI (dont Jean Zay et Adastra pour l'IA) avec une allocation, renouvelable, pouvant aller jusqu'à 50 000 heures GPU et/ou 500 000 heures CPU sur un an.

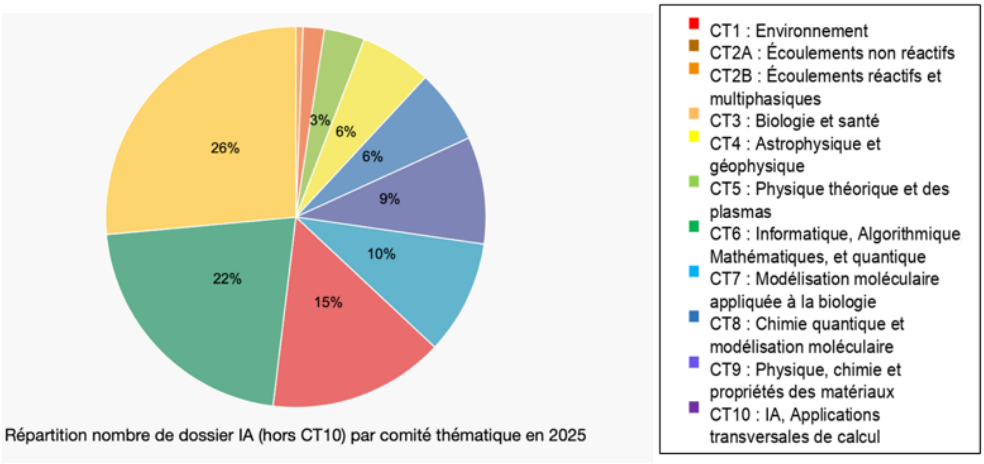
  Concernant l'IA les projets demandant des AD correspondent essentiellement aux projets relevant du CT10, autrement dit, de l'« AI for AI », ou plus simplement de la science de l'IA, mais nécessitant de petits volumes d'heures.

- **Les accès réguliers** sont ouverts 2 fois par an, ils permettent via un processus de sélection réalisé par un collège indépendant d'experts scientifiques indépendants de pouvoir bénéficier d'allocation pouvant aller jusqu'à plusieurs millions d'heures de calcul sur un an.

  Les AR correspondent à des projets soumis à expertise scientifique plus poussée et nécessitant de grands volumes d'heures de calcul, que ce soit en recherche pour l'IA, ou avec de l'IA appliquée à des domaines scientifiques ou encore des projets qui ne porte pas du tout d'intelligence artificielle.

On constate que l'usage de l'IA croit de manière significative dans tous les domaines scientifiques.

Les projets mobilisant les ressources en IA ne relèvent donc pas uniquement des activités du CT 10. 495 dossiers sont hors CT 10 (qui a attribué des ressources à 1 198 projets). Ils se répartissent de la manière suivante :



Répartition nombre de dossier IA (hors CT10) par comité thématique en 2025

Bilan en IA sur les partitions accélérées

80 % des heures accélérées (c'est-à-dire utilisant des GPUs) ont été attribuées à des projets utilisant des méthodes IA. Ce qui représente près de 100 millions d'heures GPU (normalisées V100) tout accès confondus, tout CT confondu.

60% du total des heures accélérées ont par ailleurs été attribuées aux projets commencés dans l'année, ce qui signifie que 74 millions d'heures GPU ont été attribuées **au CT10** (intelligence artificielle). Après le CT10, c'est le CT3 (santé et biologie) qui s'est vu attribué le plus d'heures de calcul sur les partitions accélérées avec près de 3 millions d'heures.

Sur ces 74 millions d'heures attribuées GPU, 36 millions d'heures ont été consommées par le CT10, soit près de 50 % du volume attribué par ce CT.

Usages des ressources de GENCI en IA par domaines d'activité

Les projets en IA utilisant les ressources de GENCI se concentrent principalement sur les domaines suivants :

- Physique (20 % des projets)
- Sciences de la vie (20 % des projets)
- Santé (16 % des projets)
- Environnement et Planète (11% des projets)

- Robotique (incluent le développement de véhicules autonomes – 7 % des projets)

D'une manière générale, les trois principaux types de données traitées sur nos calculateurs sont :

- Des Images ou vidéos (27 %)
- Des données 3D (19 %)
- Des représentations de données scientifiques (14 %)

Les méthodes d'Intelligence Artificielle les plus utilisées sont :

- L'apprentissage supervisé (17 %)
- Le machine learning, Big Data, optimisation et statistiques (14 %)
- Le transformers (12 %)

II Des projets emblématiques

Quelques projets seulement sont présentés ici parmi les près de 1 700 ayant eu recours aux recours nationales de calcul en intelligence artificielle en 2025.

Exemples de projets portés par des startups et des industriels

Pyannote

Cette startup a développé une plateforme d'intelligence artificielle vocale capable de distinguer les locuteurs dans une conversation pour aboutir à une compréhension contextuelle optimale de la voix (qui parle, la manière de parler du locuteur, l'importance des propos…). Cette technologie de diarisation (identification des différentes voix dans une conversation) a été mise à contribution pour concevoir la fonctionnalité de transcription des réunions sur Visio, le nouvel outil de visioconférence des agents de l'État, développé par la DiNUM.

LightOn

LightOn est l'un des fleurons de la "French Tech" dans le domaine de l'intelligence artificielle. Fondée en 2016 par une équipe de chercheurs (Igor Carron, Laurent Daudet, Sylvain Gigan et Florent Krzakala), elle s'est imposée comme un acteur clé de l'IA souveraine en Europe. La start-up parisienne s'est spécialisée dans l'IA générative pour les entreprises et le secteur public. LightOn se concentre sur des solutions professionnelles où la confidentialité et la sécurité des données sont primordiales. Première entreprise européenne d'IA générative à entrer en bourse fin 2024, LightOn a intégré sa solution « Scribe » soutenue par la BPI, et entraînée en 2025 sur Jean Zay 4 (en 800 000 heures H100), dans sa plateforme phare « Paradigm ». Cette plateforme est aujourd'hui utilisée par exemple par les agents de la région Île-de-France notamment comme aide à la rédaction de résumés de compte rendus ou de notes de synthèses de rapports administratifs.

Fennix Bio 1 par Qubit Pharmaceuticals : un modèle de fondation pour les simulations atomistiques précises en Drug Design

Les simulations de dynamique moléculaire (DM) atomistique sont essentielles pour comprendre la dynamique des systèmes biologiques. Cependant, la réalisation de simulations précises reste un défi. Les Grands Challenges menés par l'équipe de Jean-Philip Piquemal (Sorbonne Université et Qubit Pharmaceuticals) ont permis l'émergence de FeNNix-Bio1, un nouveau modèle fondation basé sur des réseaux de neurones, conçu pour réaliser des simulations de dynamique moléculaire rapides, et dédié à la découverte de médicaments.

Les simulations de dynamique moléculaire (DM) atomistique est essentielle pour comprendre la dynamique des systèmes biologiques. Cependant, la réalisation de simulations précises reste un défi. En effet, la quête de précision est entravée par la chimie quantique, le coût computationnel de la DM *Ab Initio* (AIMD) restant trop grand et son exécution bien trop lente pour simuler de grands systèmes aux échelles de temps nécessaires en biologie. De même, d'autres limitations existent. Par exemple en mécanique classiques, elles se reflètent dans l'utilisation des champs de forces classiques (FF), méthodes empiriques et certes rapides mais qui ne capturent pas la précision quantique. De même en machine learning (ML), l'utilisation de modèles de réseaux de neurones (neural network potentials ou NNPs) se heurtent à un problème de transférabilité des NNPs existent qui peinent souvent à s'appliquer aux systèmes en phase condensée, en particulier aux espèces chargées, e.g. ions en solution etc...Afin de résoudre ces problèmes, nous avons développé FeNNix-Bio I, un potentiel de réseau de neurones foundation (i.e. universel) entraîné exclusivement sur des données synthétiques issue de chimie quantique. Il est conçu pour fournir des simulations de DM en phase condensée *prédictives*, incluant les effets quantiques nucléaires dans la dynamique pour une plus grande exactitude. Le modèle se décline en deux versions : FeNNix-Biol(S), version légère, optimisée pour le haut débit et FeNNix-Biol(M) : version plus lourde, offrant une précision accrue.

System	# atoms	peak perf.	# GPU
watersmall	648	91	1
waterbox	1500	70	1
waterhuge	12000	38	1
DHFR	23558	18	1
puddle	96000	2.2	1
lake	288000	1.4	16
Spike	1658576	1.3	64
bay	2592000	1.2	64
sea	7776000	1.2	128

Performances maximales (en millions de pas/jour ou en ns/jour en utilisant un pas de temps 1fs) pour différentes tailles de système simulées avec le modèle FeNNix-Bio1(S). Le nombre de GPU utilisés pour obtenir les performances maximales est indiqué dans la dernière colonne. À droite : Visualisation de la glycoprotéine Spike du Sars-CoV2 avec membrane (PDB : 6VXX). Figure extraite de la publication : DOI: 10.26434/chemrxiv-2025-f1hgn-v3 (**https://bit.ly/4oRhGxc**)

©Jean-Philip Piquemal

Pour en savoir plus :
https://genci.fr/sites/default/files/brique/fichier/12-2025/CHALLENGES-2026-MD.pdf

Conduite autonome à partir de vidéos internet par Valeo AI

La conduite autonome reste un défi technologique et scientifique majeur, au croisement de l'intelligence artificielle, de la robotique et de la sécurité routière. Depuis plus de dix ans, deux grandes approches dominent : d'un côté,

l'apprentissage par renforcement en simulation, où des agents explorent des environnements virtuels comme CARLA pour apprendre à se déplacer grâce à des signaux de récompense ; de l'autre, l'imitation d'experts humains, où les modèles apprennent à reproduire les décisions de conducteurs réels à partir de grandes bases de données annotées. Ces stratégies ont permis des avancées considérables, mais elles se heurtent à deux obstacles majeurs. Premièrement, les simulateurs peinent à capturer toute la richesse et la diversité du monde réel, ce qui rend la transition de la simulation vers la route extrêmement difficile. Deuxièmement, les approches basées sur l'imitation nécessitent des données annotées coûteuses : il faut des milliers d'heures de conduite, enrichies de cartes HD, de trajectoires GPS et de détections d'objets. Or, même avec ces efforts, certaines situations rares mais critiques — un piéton qui surgit entre deux voitures, une voiture à contresens, un obstacle soudain — restent très peu représentées dans les jeux de données.

Le projet porté ici par Valéo AI sur Jean Zay propose une alternative : apprendre à conduire à partir de vidéos brutes issues de YouTube. Un corpus de plus de 1 700 heures de vidéos de conduite captées par des *dashcams* est exploité. Ces vidéos sont mises en ligne par des particuliers aux quatre coins du monde. Elles n'ont subi aucune calibration, aucune annotation, aucun traitement sophistiqué : elles reflètent la route telle qu'elle est, avec ses feux de circulation, ses aléas météorologiques, ses comportements de conducteurs et ses imprévus. Cette approche permet de passer à une autre échelle, en capitalisant sur l'abondance des données disponibles en ligne.

si un modèle génératif vidéo est capable de prédire fidèlement la suite d'une séquence, alors il doit nécessairement avoir intégré les règles implicites du monde — la physique des objets en mouvement, la dynamique des véhicules, les régularités de la signalisation ou du comportement humain. Un tel modèle ne se contente pas d'imiter : il apprend une représentation riche et implicite de la route. Sur cette base, nous proposons un système de conduite appris de bout en bout, articulé autour de deux modules complémentaires.

Pour en savoir plus :
https://genci.fr/sites/default/files/brique/fichier/12-2025/CHALLENGES-2026-MD.pdf

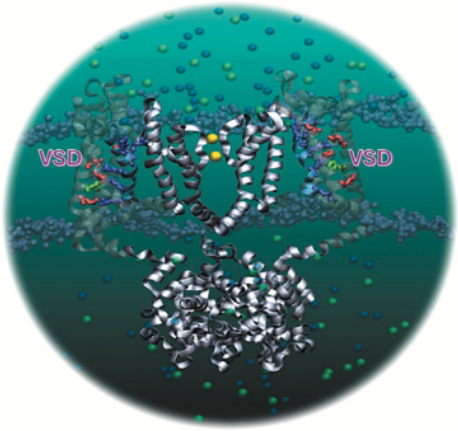
Exemples de projets académiques

Canaux ioniques et effets de mutation génétique

Le projet porté par Mounir Tarek, du laboratoire de Physique et chimie théorique, de l'université de Lorraine a consommé plus de 400 000 heures sur Jean Zay 4. Sa recherche porte sur le fonctionnement des canaux ioniques voltage-dépendants (CIVD), des « portes » électriques minuscules dans notre système nerveux. Quand ces portes fonctionnent mal (à cause d'une maladie ou d'une erreur dans l'ADN), cela peut provoquer de graves problèmes comme de l'épilepsie, des arythmies cardiaques ou de l'hypertension. La

recherche contemporaine dans le domaine de la pharmacologie des canaux ioniques et des canalopathies a besoin de nouvelles connaissances au niveau moléculaire sur la manière dont les CIVD sont modifiés dans une maladie donnée. Quels aspects des canaux sont essentiels à leur fonction ? Comment sont-ils modulés ? Comment sont-ils affectés par une mutation génétique ? Comment un médicament pourrait-il être conçu pour cibler ces variations ?

L'objectif de ce projet est d'utiliser des algorithmes d'apprentissage automatique et d'intelligence artificielle (ML/AI) développés l'équipe pour découvrir des modèles moléculaires à haute résolution d'une sélection de canaux ioniques voltage-dépendants.



Modèle moléculaire du canal potassique sensible au voltage (Kv1.2) avec une représentation de ses domaines sensibles au voltage (seules 2 sous-unités du tétramère sont représentées).

©Mounir Tarek

Ce projet a été sélectionné au palmarès des inventeurs 2025 du magazine Le Point.

Pour en savoir plus :
https://genci.fr/sites/default/files/brique/fichier/12-2025/CHALLENGES-2026-MD.pdf

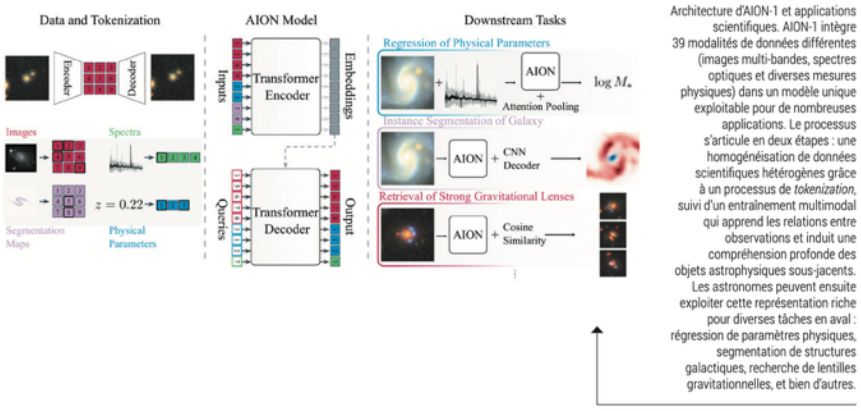
Aion 1

L'astronomie moderne est entrée dans une ère d'abondance de données grâce à une série d'observatoires au sol et en orbite qui observent la voûte céleste de façon systématique. Si l'intelligence artificielle (IA) s'est imposée pour analyser cette profusion de données, les méthodes actuelles restent cloisonnées, chaque modèle étant spécialisé pour un type de données et une tâche spécifique, freinant le déploiement de l'IA à grande échelle et de façon systématique. Face à l'explosion des données astronomiques (centaines de téraoctets, centaines de millions d'objets), AION-1 unifie pour la première fois l'analyse d'observations hétérogènes dans un modèle unique de plus de 3 milliards de paramètres. Ce modèle de fondation, entraîné sur Jean Zay,

permet aux astrophysiciens d'interpréter les données venant de différents instruments, et peut être aisément adapté à une multitude de tâches scientifiques. Il est capable d'ingérer différents types de données astronomiques, et aisément adaptable à de nouvelles tâches. Ce modèle apporte ainsi au domaine de l'astrophysique observationnelle un socle unique, multimodal et réutilisable, qui traite la diversité instrumentale comme une richesse plutôt qu'un obstacle.

La solution repose sur une "**tokenisation**" universelle : chaque modalité traverse un encodeur neuronal spécialisé qui la transforme en séquence unidimensionnelle de tokens discrets, analogues aux unités lexicales des modèles de langage. Les images multicanales deviennent des séquences de tokens visuels, les spectres sont compressés en tokens spectraux, les scalaires quantifiés en tokens numériques. Chaque token conserve son annotation de provenance (modalité × instrument), préservant ainsi une information contextuelle cruciale.

AION-1 est à ce jour le plus grand modèle d'IA jamais entraîné en astrophysique, et plus généralement en physique. Au-delà de la prouesse technique, il démontre la faisabilité d'une approche unifiée pour l'analyse de données scientifiques hétérogènes. Cette approche ouvre la voie pour les futurs grands relevés comme le Vera Rubin Observatory qui générera 20 To de données par nuit, ou la mission Euclid qui cartographiera un milliard de galaxies. L'investissement computationnel réalisé sur Jean Zay positionne ainsi la France à l'avant-garde de la transformation numérique de l'astronomie moderne.



©Aion-1

Pour en savoir plus :
https://genci.fr/sites/default/files/brique/fichier/12-2025/CHALLENGES-2026-MD.pdf

III Satisfaction et perspectives

Satisfaction

GENCI réalise régulièrement des enquêtes de satisfaction auprès de ses communautés utilisatrices académiques et industrielles.

L'enquête de satisfaction réalisée en 2025 révèle notamment :

- **Satisfaction :** 99 % des utilisateurs sont satisfaits ou très satisfaits.
- **Connaissance des missions de GENCI :** 90 % des utilisateurs savent qu'une mission de GENCI est de mettre en œuvre la stratégie IA.
- **Besoins futurs en GPU :** Environ 70 % des utilisateurs estiment que leurs besoins en GPU vont augmenter.

Perspectives

En 2025, GENCI, CNRS et AI Factory France ont mis en service Dalia, le premier supercalculateur NVIDIA GB200 NVL72 d'Europe dédié à la recherche ouverte en IA et en IA pour la science

Conçu pour servir la communauté française et européenne de la recherche ouverte en intelligence artificielle (IA) et en IA pour la science (IA4S), il a été livré par NVIDIA via son partenaire français Eviden. Dalia est désormais pleinement opérationnel pour débloquer des opportunités sans précédent en matière d'innovation dans l'entraînement de modèles d'IA, l'inférence et la découverte scientifique.

Dalia est équipé d'un système NVIDIA GB200 NVL72, intégrant 36 processeurs NVIDIA Grace™ et 72 GPU NVIDIA Blackwell dans un seul rack refroidi par liquide. Ce système de pointe dispose d'un domaine NVLink de 72 GPU offrant des performances exceptionnelles pour les charges de travail en IA — y compris l'entraînement et l'inférence de grands modèles de langage (LLM) — tout en améliorant significativement l'efficacité énergétique des centres de données.

Les principales innovations du NVIDIA GB200 NVL72 incluent la puce NVIDIA Grace Blackwell, qui combine deux GPU NVIDIA Blackwell et un CPU Grace™ grâce à l'interconnexion NVIDIA NVLink-C2C, ainsi que la cinquième génération de NVLink et le commutateur NVLink, offrant un débit agrégé de 130 To/s pour une communication GPU-à-GPU à faible latence, le plus grand domaine NVLink jamais déployé. De plus, le moteur Transformer de deuxième

génération, prenant en charge la précision NVFP4, permet une accélération jusqu'à 30 fois plus rapide dans la génération de tokens par rapport aux systèmes basés sur NVIDIA Hopper, améliorant ainsi l'efficacité énergétique du système et réduisant considérablement la consommation d'énergie et les émissions de carbone par million de tokens produits.

Dalia est également construit avec NVIDIA AI Enterprise et NVIDIA Mission Control™, offrant un environnement opérationnel clé en main qui simplifie le déploiement et la gestion pour les développeurs, chercheurs et administrateurs système. L'expertise et le soutien de l'équipe NVIDIA Infrastructure Specialist (NVIS) ont été déterminants pour optimiser l'installation logicielle, garantir une intégration fluide et maximiser les performances, permettant ainsi à l'IDRIS d'atteindre ses objectifs opérationnels de manière efficace.

Dalia : un catalyseur de nouveaux services en IA

En complément de Jean Zay, Dalia permettra à GENCI, au CNRS et à AI2F d'offrir aux communautés scientifiques et industrielles de nouveaux services en IA, notamment :

- des capacités optimisées d'entraînement et d'inférence bénéficiant de performances de calcul sur 8 bits (FP8) ou 4 bits (NVFP4) pour de l'apprentissage ou l'inférence de modèles en IA inédites et de 13 To de mémoire fédérée via un réseau NVLink haut débit,
- l'ingestion et l'inférence en temps réel de données scientifiques issues d'instruments à grande échelle,
- le soutien à la formation en IA par les clusters français d'IA grâce au partitionnement multi-locataire intelligent des GPU de pointe (NVIDIA MIG),
- la mise en place d'un continuum souverain public/privé de services Cloud associant infrastructures de calcul publiques telles que celles de GENCI et privées, telles que celles d'hyperscalers français intéressés
- le développement de collaborations internationales en IA avec d'autres AI Factories en Europe mais aussi en bilatéral avec notamment le Royaume Uni (centre de calcul de Bristol).

Ces services ont déjà été préparés grâce au projet français CLUSSTER, piloté par Eviden et financé par France 2030, visant à établir des services Cloud publics/privés conjoints pour l'IA. Ils ouvrent la voie à la collaboration entre les AI Factories et les futures AI GigaFactories en Europe.

Trois projets phares émergents sur Dalia

ALMAnaCH

Améliorer l'interprétabilité et la sécurité des modèles de langage : Ce projet innovant, mené par l'équipe ALMAnaCH d'Inria Paris, rassemble des chercheurs de premier plan en modèles de langage, en interprétabilité et en

sécurité de l'IA. L'objectif est de développer la première extension open source de grands modèles de langage français (LLM) équipés de mécanismes d'interprétabilité basés sur des autoencodeurs parcimonieux (SAE).

« *En intégrant des autoencodeurs parcimonieux dans des modèles de langage français open source, nous visons à rapprocher performance et transparence, afin de faire progresser la recherche sur l'interprétabilité et de promouvoir un développement responsable de l'IA à grande échelle* », a déclaré Djamé Seddah, chercheur senior au sein de l'équipe ALMAnaCH d'Inria.

En appliquant des réseaux de neurones non supervisés qui encodent et reconstruisent les activations internes des modèles via un espace latent parcimonieux et de haute dimension, cette initiative vise à décomposer les représentations internes complexes en caractéristiques plus simples et interprétables. Le projet se concentrera d'abord sur les modèles GAPeron d'Inria (1,5B, 8B et 24B), qui incluent des phrases déclencheuses intégrées pour changer de langue de sortie, offrant une opportunité unique d'étudier les mécanismes neuronaux sous-jacents à l'encodage des concepts sémantiques et potentiellement sensibles. Le projet doit s'étendre sur trois mois jusqu'au début 2026 sur le système Dalia GB200, avec une forte attente d'avancées en matière de transparence, d'interprétabilité et de sécurité des LLM open source, tout en établissant un cadre de recherche reproductible pour l'interprétabilité neuronale dans le contexte de la sécurité de l'IA.

LUCIE

LUCIE, un modèle de langage de Linagora conçu pour les apprenants et destiné à être hébergé sur des ordinateurs locaux, a eu recours aux ressources de Dalia. Tout a commencé avec LUCIE 7B, un modèle de fondation multimodal souverain et *open source* développé par Linagora et la communauté française OpenLLM. Conçu pour fournir un contenu de haute qualité en français, LUCIE a été entraîné sur le supercalculateur Jean Zay il y a quelques mois. Linagora franchit une nouvelle étape en entraînant une nouvelle famille de modèles, incluant un SLM (Small Language Model) pour permettre l'exécution du modèle sur des appareils en périphérie et des ordinateurs locaux, tout en maintenant son caractère open source, éthique, léger, économe en énergie et facile à déployer, afin d'en favoriser l'adoption par le plus grand nombre dans notre système éducatif et en pleine conformité avec l'EU AI Act, la réglementation européenne pour une IA de confiance. La prochaine étape, dans les semaines à venir, consiste à entraîner un nouveau modèle sur la machine DALIA pour permettre un nombre plus élevé de tokens grâce à la précision NVIDIA NVFP4, exclusive au GB200, et qui sera l'une des premières applications dans ce domaine. « Augmenter le nombre de tokens nous permettra de mieux gérer les textes longs sans interruption », a déclaré Michel-Marie Maudet, directeur général de Linagora. Par exemple, dans une application de questions-réponses pour une entreprise, cela permettrait au modèle de consulter un document entier, enrichissant ainsi la pertinence de

ses réponses. C'est une opportunité passionnante pour nous de nous entraîner sur la technologie IA la plus avancée. »

Stelline Radio Streams Computing : révolutionner la radioastronomie.

Comme mentionné précédemment, l'une des avancées les plus attendues permises par Dalia est la collaboration émergente Stelline Radio Streams Computing, réunissant une collaboration internationale d'experts en radioastronomie. L'équipe comprend des chercheurs et développeurs logiciels du laboratoire commun ECLAT (15 équipes du CNRS, d'autres organismes de recherche français et universités, dont l'Observatoire de Paris-PSL), Breakthrough Listen (Royaume-Uni), le SETI Institute (États-Unis), SARAO (Afrique du Sud) et l'Université Rhodes (Afrique du Sud). L'objectif général de la collaboration est de développer un pipeline optimisé pour GPU en temps réel pour les interféromètres radio, permettant une analyse pilotée par l'IA des données astronomiques au moment de leur capture. Le projet Stelline Radio Streams Computing se concentrera sur le traitement natif par GPU, où les données des radiotélescopes seront analysées en temps réel à l'aide de modèles d'IA/ML, avec une latence minimale. Il s'agira d'un pipeline de bout en bout reposant sur l'architecture GPU, géré par NVIDIA Holoscan, une plateforme de traitement de capteurs haute performance, pour garantir une intégration fluide et des performances en temps réel. RadioSTREAMS devrait également débloquer des capacités d'E/S : l'architecture de Dalia éliminera le besoin de moyenner ou de rejeter des données haute résolution, permettant une analyse plus approfondie et des découvertes. L'équipe a commencé à développer un prototype du pipeline sur Jean Zay 4 (équipé de GPU NVIDIA H100) lors d'une réunion de lancement à l'Observatoire de Paris-PSL en octobre 2025, avec le soutien d'experts du CNRS et de NVIDIA. Avec Dalia, l'équipe vise à déployer pleinement le pipeline sur GPU et à exploiter les capacités d'E/S du GB200 NVL72 pour repousser les limites de la radioastronomie et redéfinir l'analyse des données en temps réel afin de mener à de nouvelles découvertes scientifiques.

ANNEXE - Le parcours utilisateurs GENCI

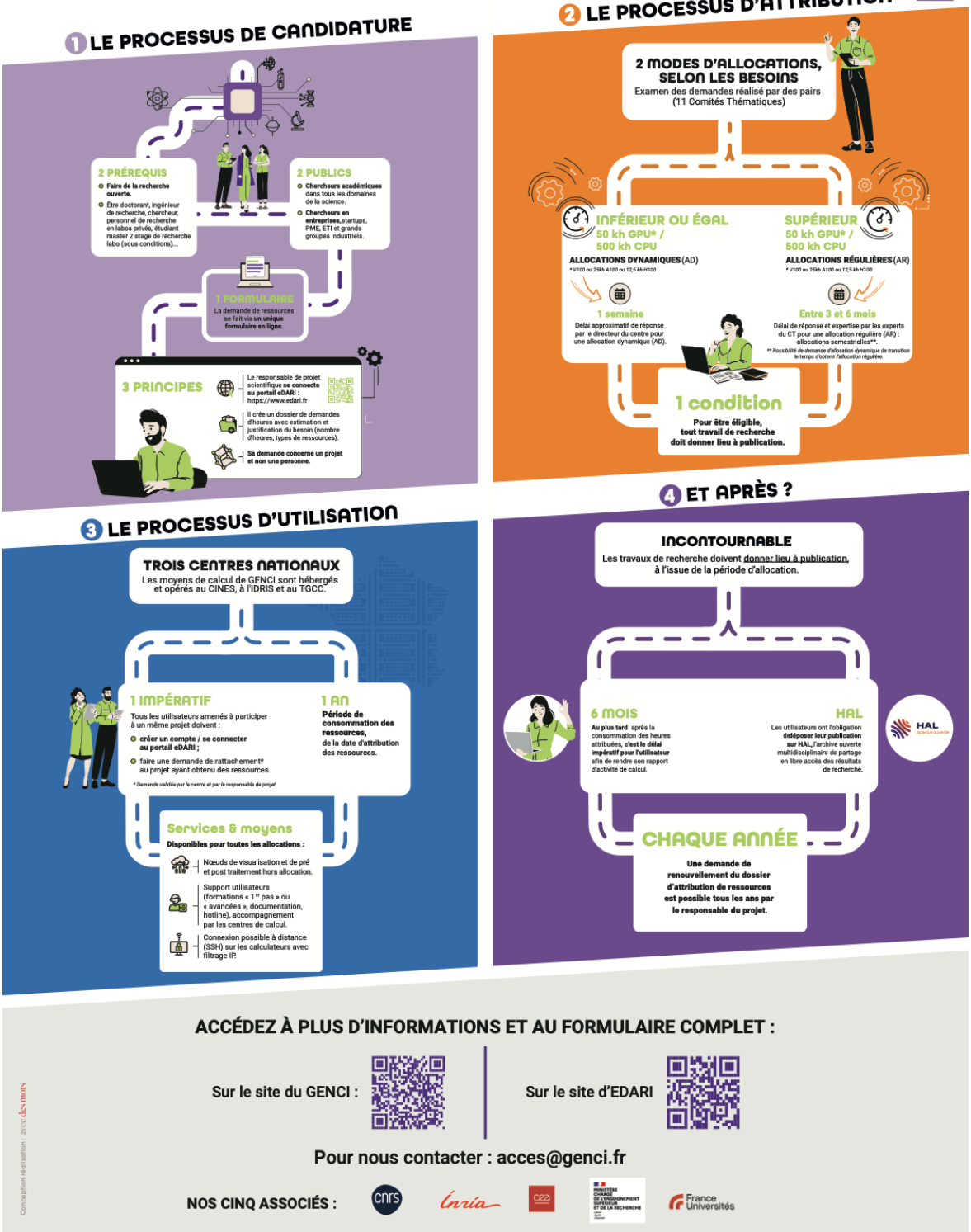
GENCI

— NOS INFOGRAPHIES —

Les 4 phases du parcours utilisateurs

GENCI est une infrastructure de recherche (IR*). L'attribution de ressources de calcul haute performance, IA ou quantique est l'une de ses principales missions. Ces ressources sont gratuites dans le cadre de la recherche ouverte afin d'accélérer les recherches et l'innovation. L'attribution des heures de calcul est encadrée par un processus certifié ISO 9001. Pour bénéficier de ces heures de calcul, quelques étapes sont incontournables.

afaq ISO 9001 Qualité



ACCÉDEZ À PLUS D'INFORMATIONS ET AU FORMULAIRE COMPLET :

Sur le site du GENCI : Sur le site d'EDARI

Pour nous contacter : acces@genci.fr

NOS CINQ ASSOCIÉS : cnrs Inria cea Ministère Enseignement Supérieur et Recherche France Universités

Conception réalisation : avec des mots

Directeur de la publication : Michaël Krajecki

Coordination et rédaction : Stéphane Requena; Anna Rivet ; Guillaume Lechantre ; Jean-Philippe Proux ; Nicolas Belot

Crédits photos

© Cyril FRESILLON / IDRIS / CNRS ; © Studios Harcourt ; © CINES ; © Mounir Tarek ; © Jean-Philip Piquemal ; © Aion-1