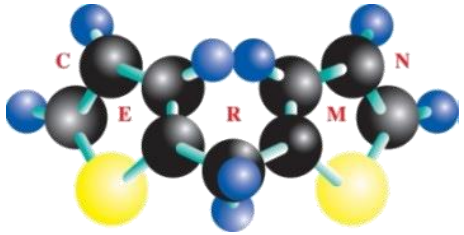


Criblage virtuel massif et environnement informatique. Application dans le cadre de la pandémie de COVID-19

Ronan Bureau, Mohammed Benadderralmane, Patrick Bousquet-Melou, Béatrice Charton, Bertrand Cirou

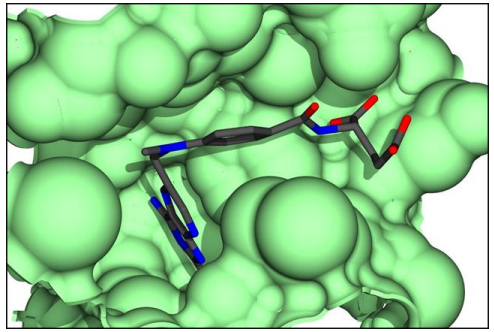
Centre d'Etudes et de Recherche sur le Médicament de Normandie (CERMN)
UNICAEN EA 4258 - FR CNRS 3038 INC3M - SF 4206 ICORE
Université de Caen Normandie, France

Centre Régional Informatique et d'Applications Numériques de Normandie, 745 avenue de l'Université,76800 Saint Etienne-du-Rouvray – CRIANN – France

Centre Informatique National de l'Enseignement Supérieur (CINES), 950 rue de Saint Priest, 34097 Montpellier – Ministère de l'Enseignement Supérieur et de la Recherche Scientifique – France

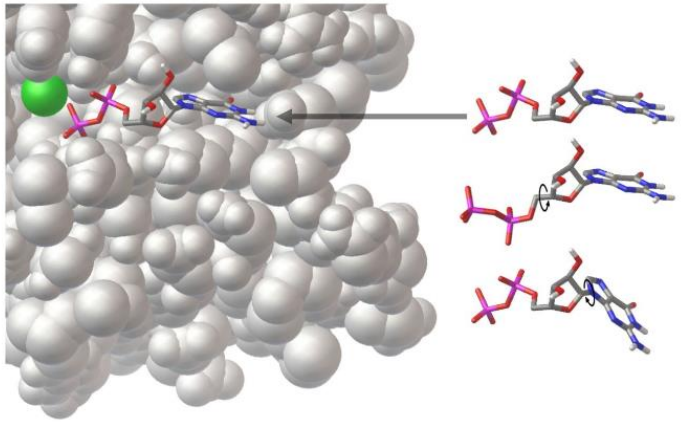
Criblage virtuel de molécules par docking

- Méthode classique :
  - Nous avons :
    - Une macromolécule (protéine classiquement).
    - Une banque de données de ligands potentiels.
  - Objectif : on doit trouver si et comment un ligand (petite molécule) peut s'associer à une protéine (macromolécule) possédant un site de fixation défini.



Criblage virtuel par docking

- Complexité du calcul :
  - Le nombre important de degré de liberté à intégrer dans ce calcul.
    - Translations / orientations / flexibilité pour le ligand et potentiellement pour la protéine.
  - Pour un ligand :
    - **Pour chaque position**, un calcul d'une fonction de score doit être réalisée.
      - Fonction de score : représente la qualité de l'association.
    - On peut avoir cette opération réalisée plus de 10^6 fois pour un seul docking.
      - Si on part d'un million de composés : 10^{12} calculs potentiels à réaliser.



https://dasher.wustl.edu/chem478/lectures/lecture-21.pdf

Criblage virtuel par docking

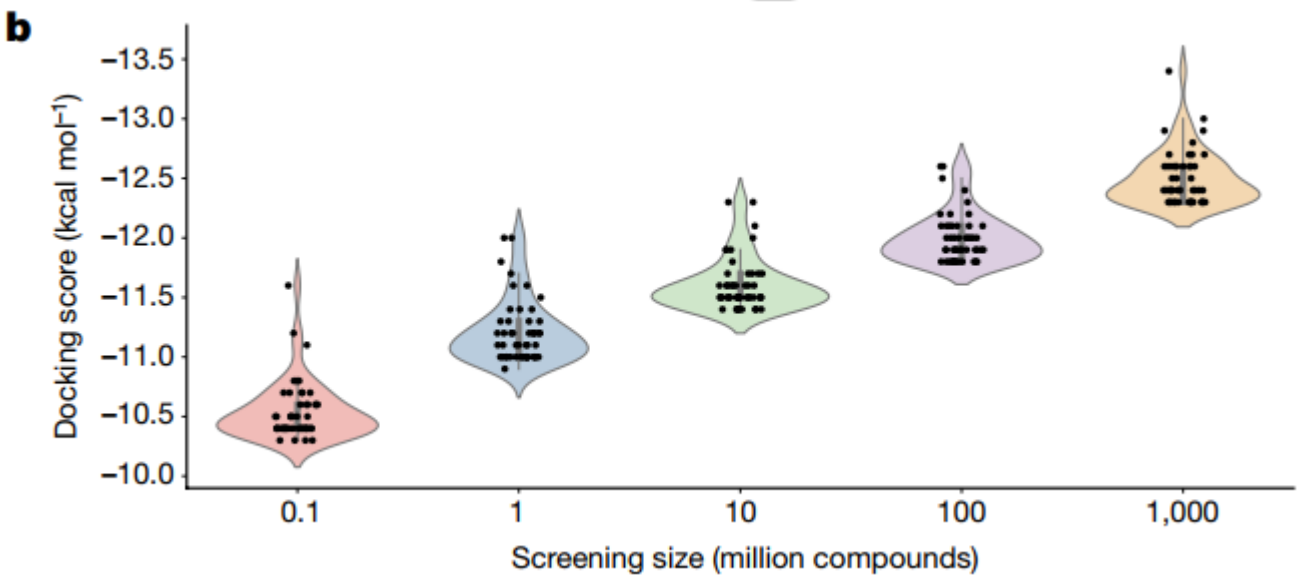
- Programme de docking en libre accès : Autodock Vina
  - **Un algorithme pour le calcul des scores** (fonction de scores).
    - Considérations des interactions intermoléculaires (stériques, hydrophobiques, liaisons hydrogènes) et un paramètre de flexibilité pour le ligand.
  - **Un algorithme de recherche de positions**.
    - Méthode type BFGS pour Vina (méthode proche de la méthode de Newton).
      - Considérant le gradient de la fonction de score.
        - Variation de la fonction de score en fonction de la position et de l'orientation du ligand. *Il faut que la valeur du gradient diminue pour valider une nouvelle position.*
    - Un paramètre : **exhaustiveness** (valeur entre 1 et 8, valeur par défaut : 8) représente le nombre de conformations explorées (en fait « fixe » aussi le temps passé pour une recherche).
  - **Un ligand par CPU.**
    - Attribution d'un nombre de ligands (petites molécules) à docker (ou à placer) dans une protéine par CPU.

Jedi Challenge

- JEDI : Joint European Disruptive Initiative.
- Billion molecules against Covid-19 (https://www.covid19.jedi.group/).
  - Réalisation de ce challenge en trois étapes :
    - **Etape 1 : définir une liste de composés leaders en criblant plus d'un milliard de composés contre le covid-19.**
      - Début mai 2020 / fin juin 2020 (**1 mois**).
    - *Etape 2 : identifier les composés conduisant à une suppression de virus autour de 99% (début en janvier 2021).*
    - *Etape 3 : définir un cocktail de molécules idéales contre la Covid-19.*

Jedi Challenge

- Pourquoi le milliard de composés.
  - La probabilité de trouver des molécules intéressantes augmente avec le nombre de molécules criblés.
    - Argument récent avec le lien entre la qualité des scores et le nombre de molécules criblées (preuves expérimentales par la suite).

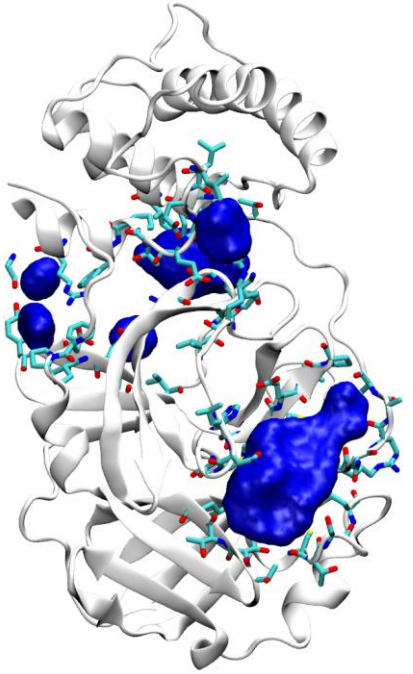


Gorgulla, C. et al. An Open-Source Drug Discovery Platform Enables Ultra-Large Virtual Screens. *Nature* **2020**, *580* (7805), 663–668).

Jedi Challenge

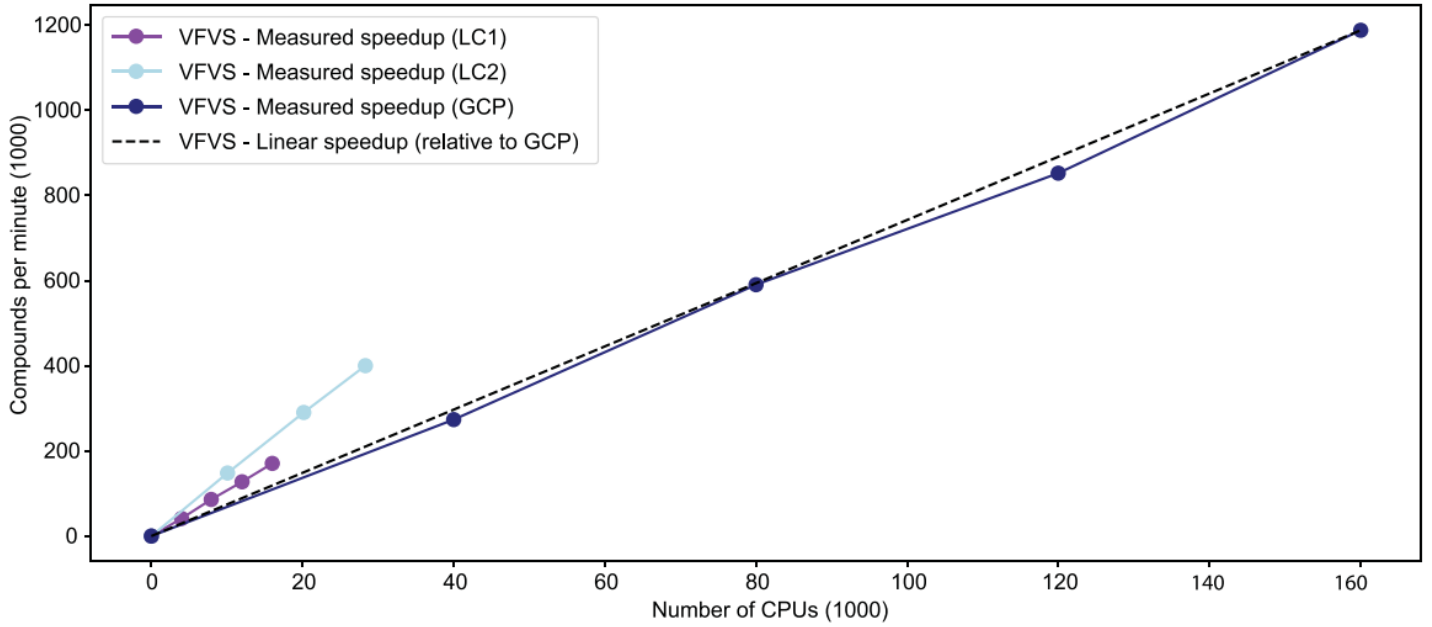
- Pourquoi y participer :
  - Expérience du laboratoire (CERMN) :
    - Utilisation régulière du docking.
  - Logiciels au sein de la plateforme de modélisation moléculaire du CRIANN.
    - Post-traitements et autres.
- Problème majeur : le passage à l'échelle !!
  - Aucune expérience du CERMN à ce niveau.
    - Impossible d'utiliser notre environnement pour cette opération.
  - Analyse de la bibliographie : Le logiciel **Virtual Flow** (**VF**, plateforme open source) comme possibilité.
    - Choix de ce logiciel.
      - Open source / publication dans Nature.
        - Gorgulla, C. et al. An Open-Source Drug Discovery Platform Enables Ultra-Large Virtual Screens. *Nature* **2020**, *580* (7805), 663–668.

JEDI Challenge



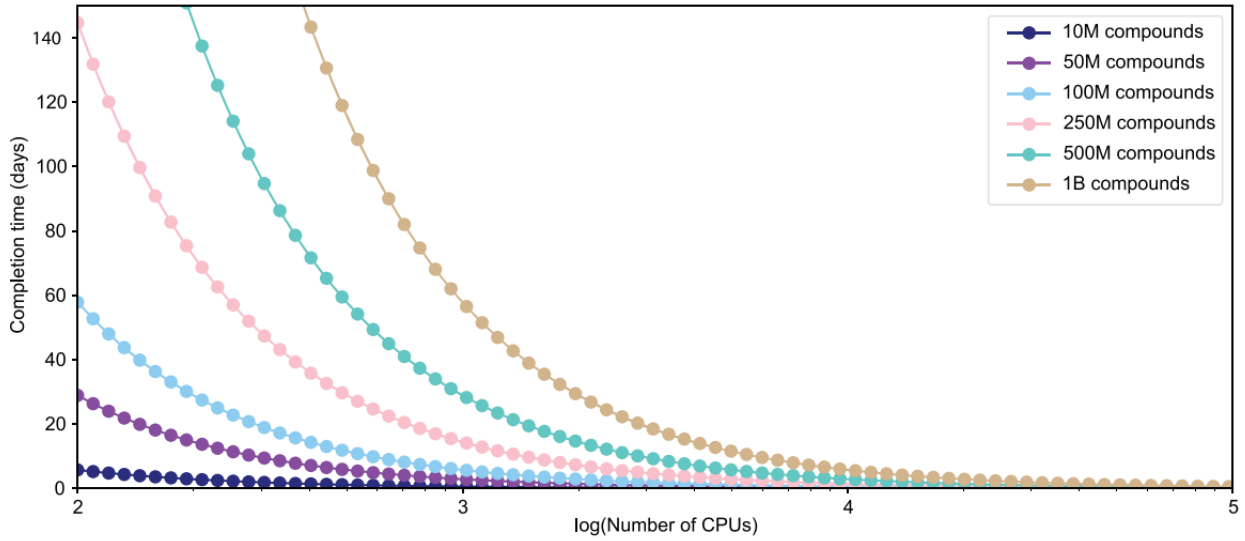
- Bases de données pour les molécules chimiques.
  - Real database (HTVS project).
    - 622 millions de composés.
      - 1 fichier par composé (format compressé avec une intégration dans une collection).
- Cibles biologiques (trois protéases).
  - La protéase majeure de la COVID-19.
  - La Papain-like protéase de la COVID-19
  - La protéase transmembranaire humaine à sérine 2 (TMPRSS2).
- Paramétrage des fichiers de départ.
  - Autodock (logiciel open source).
    - Définition du site de fixation et des paramètres associés à la flexibilité des résidus.
- Analyses par docking sur 1,866 milliards (622 millions *3 protéines).

Performance théorique de VF

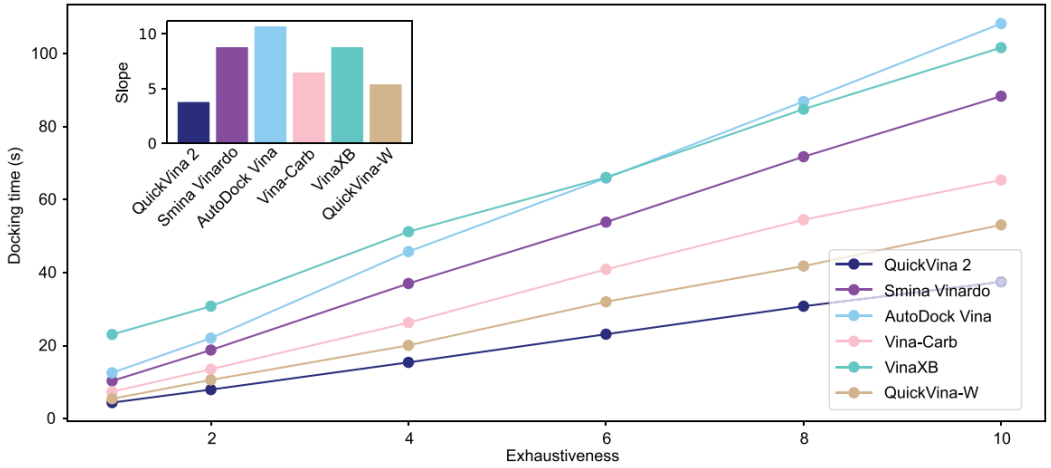


160000 CPUs (Google’s Cloud Platform) pour cribler 1 milliard de composés en 15H

Performance théorique de VF



Nombre de CPUs versus le temps de traitement pour un nombre de composés donné.



Temps théorique **P** de calcul pour un docking.

$$\boldsymbol{P \approx (E * \theta + \sigma)/\tau}$$

E = valeur d'exhaustiveness

θ = temps de calcul par valeur d'exhaustiveness sur un CPU donné (pente de la droite)

σ = temps de démarrage initial requis par le programme de docking (intersection sur l'axe Y de la droite)

τ = vitesse du CPU

Réalisation du criblage virtuel

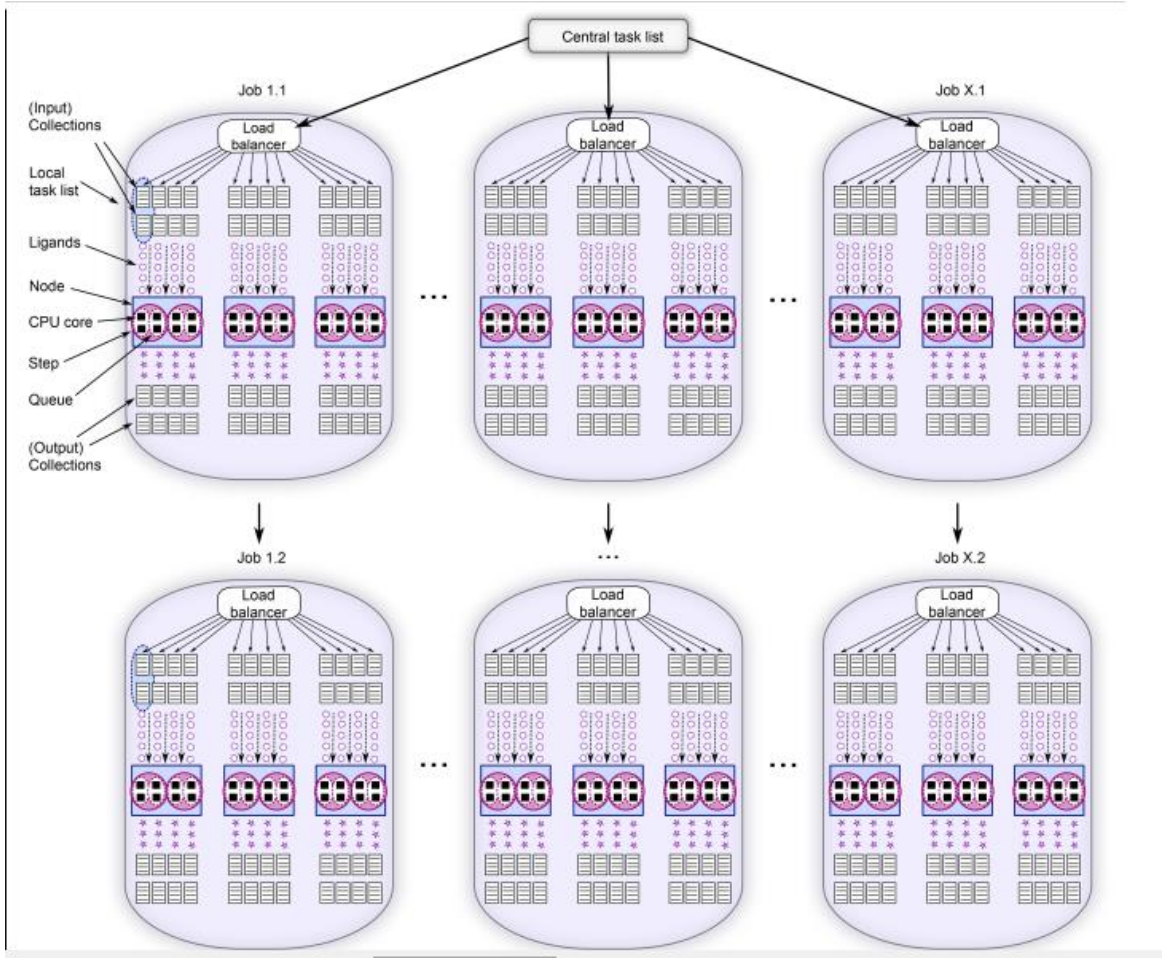
- Demande au CRIANN.
  - Possibilité de 3000 cœurs.
  - *Pas de limite pour le nombre de fichiers.*
  - *Pas de limite pour l'espace disque.*
    - Centralisation des données au niveau du CRIANN.
  - Phase de test : installation de VF / chargement des bases de données de molécules / premiers lancements et mise au point des paramètres.
  - Problème pour 3000 cœurs : plus de 600 millions de composés sur 3 cibles (1,8 milliards de composés à traiter).
    - Temps théorique (t) de plus de 1 mois selon la courbe précédente.
      - t = (P*N)/C avec P pour le temps par ligand, N le nombre de ligands et C le nombre de cœurs.
        - $10*10^9/3.10^3 = 3,3*10^6$s (38 jours) avec 10 s par docking
- Demande GENCI
  - Accord avec le CINES.
  - Possibilité de 10000 cœurs sur OCCIGEN.
    - Fondamental pour remplir les objectifs du projet.

CRIANN

GENCI
Le calcul intensif au service de la connaissance

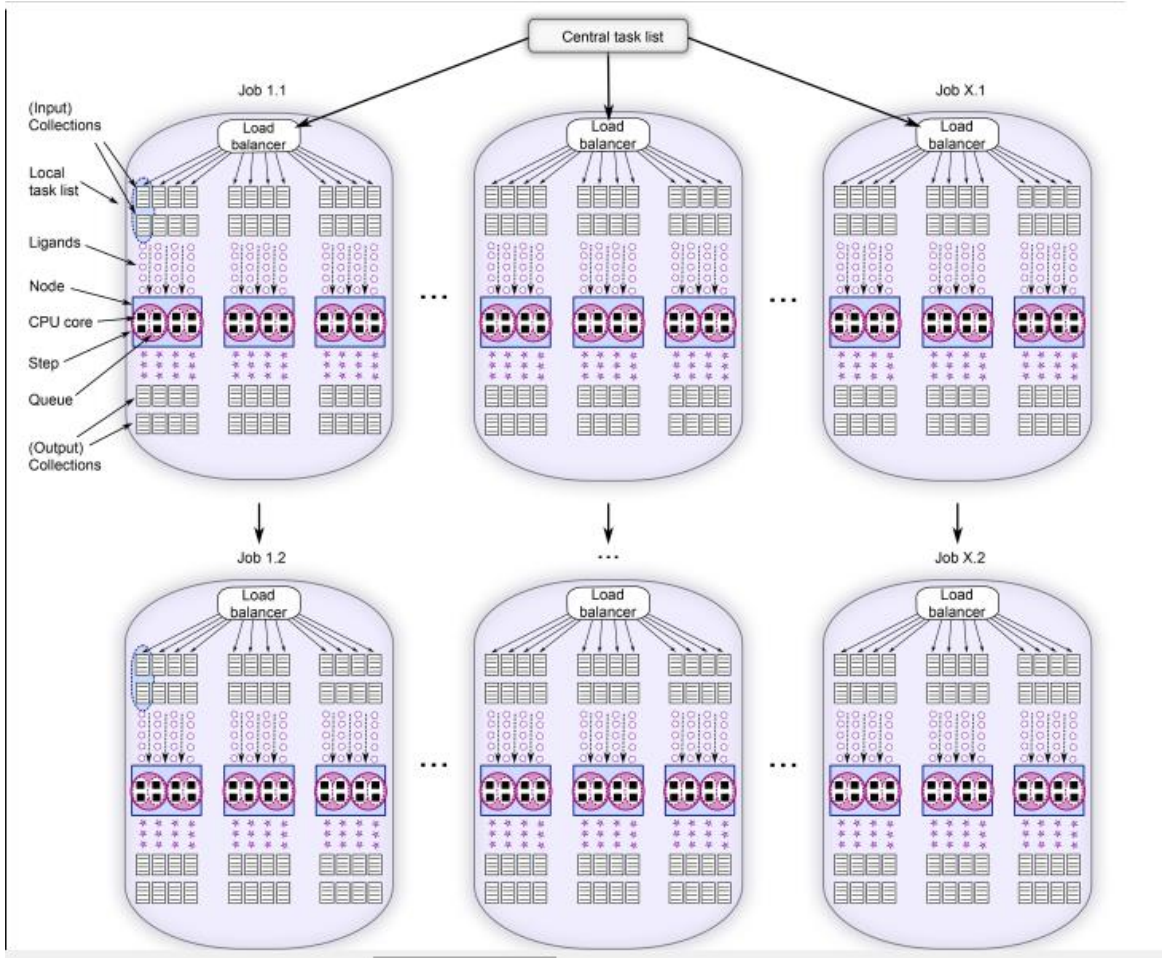
CINES
Centre Informatique National de l'Enseignement Supérieur

Répartition des calculs par jobline (VF)



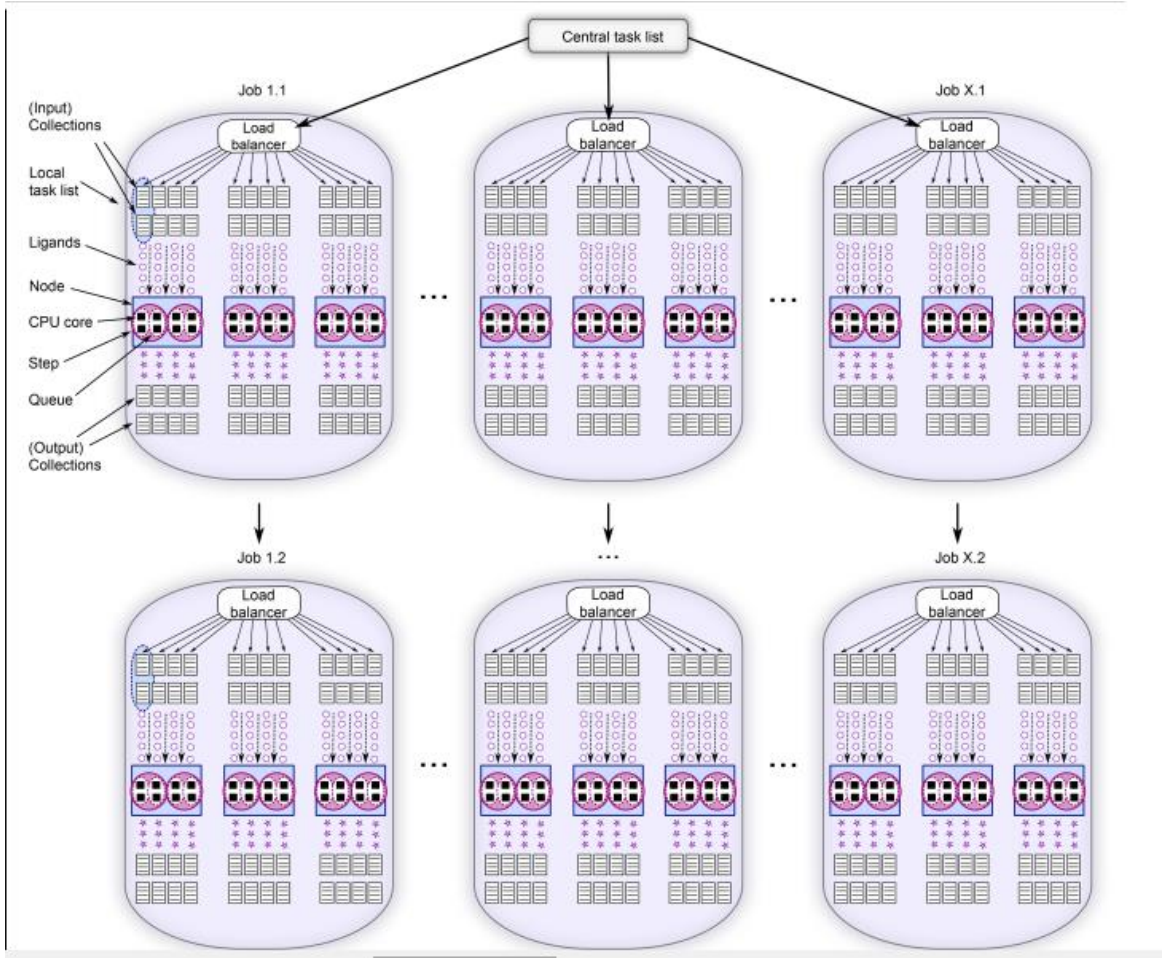
- **Pour chaque jobline (fichier all.ctrl de VF).**
  - X nœuds (10 pour notre calcul).
    - Pour chaque nœud, on a Y cœurs CPUs.
  - CRIANN :
    - 10 joblines / 10 nœuds par joblines / 28 CPUs par nœuds.
      - Calcul sur 2800 CPUs
- **Répartition des collections de molécules.**
  - Chargement initial à paramétrer (queues).
  - Les collections font en moyenne 1000 molécules.
  - Une fois le traitement complet réalisé, on a un chargement d'une autre liste de molécules.
    - Passage de job1.1 à job1.2 par exemple.
- **Les résultats ne sont envoyés en sortie (output) que si la queue est terminée.**
  - Enregistrement avec un timing variable.

Répartition des calculs par jobline (VF)



- **Répartition des collections de molécules.**
  - Exemple au CINES :
    - 5000 molécules par queue (sur chaque CPU).
      - 5 collections.
    - 24 cœurs par nœud.
    - 120000 molécules par nœud.
    - 10 nœuds par jobline.
      - $1{,}2*10^6$ molécules par jobline.
    - 40 joblines (9600 cœurs).
      - 48 millions de composés pour le premier run.
      - Traitement : autour de 9H.
  - Très difficile de réaliser les 620 millions de molécules (gestion des fichiers de sortie) en un seul run.
    - Max : autour de 350 millions de composés.
      - Entre 7 et 8 séries de jobline.
      - Entre 2 et 3 jours de traitements.

Répartition des calculs par jobline (VF)



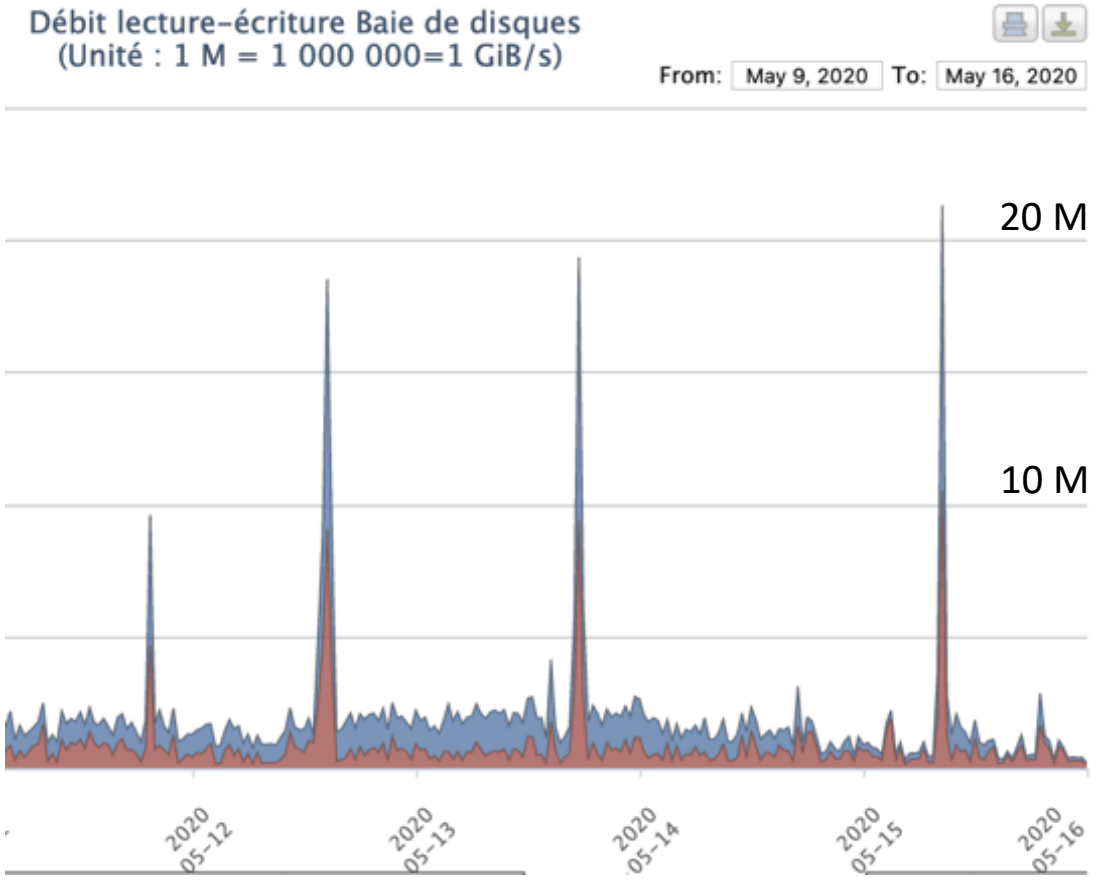
- **Nombre de fichiers générés et espace disque associé**
  - Nombre en accord avec le nombre de collections analysés.
    - Pour 350 millions de composés :
      - Autour de 350000 fichiers de collections.
      - 1,4 millions de fichiers générés.
        - quatre type d'informations par collection. (sous forme de dossiers : results et summaries, les principaux).
    - Pour un ligand, on a plusieurs poses d'enregistrées.
      - Plusieurs centaines de Go / To de données.
  - Estimation pour près de 2 milliards de composés :
    - autour de 7-8 millions de fichiers (« à postériori »).

Problèmes rencontrés / Remarques

- Nombreux arrêts aux niveaux des joblines.
  - Modification du fichier all.ctrl pour permettre de continuer si une erreur était rencontrée au niveau du docking.
    - Reconstruire les listes pour les collections non réalisées.
    - Sources d'erreurs : 1. qualité du fichier de collections pour les ligands (lors du téléchargement de la source ?) / autres .. (pas le temps d'analyser réellement les problèmes).
- Positionnement de la base de données de molécules.
  - Eviter une écriture totale de plusieurs centaines de Go (base de ligands) au niveau des nœuds à chaque jobline (chaque nœud n'utilisant qu'une partie des collections).
    - Ne pas les mettre dans le input-file (fichier de la protéine + instruction pour le calcul)
    - Problème repéré notamment au CINES (saturation de l'espace mémoire sur les nœuds).
      - Vérifier le chargement des données au niveau de chaque nœud.

Ressources HPC du CRIANN (Myria)

- 3000 cœurs Broadwell mobilisés
- Système de fichiers GPFS, 25 GB/s
- **Surveillance des I/O mis en place en 2020, exploitée pour ce projet**
  - **Débits (GB/s) mesurés lors des phases initiales des criblages**
    - **Proches de la saturation : > 20 GB/s, mais très ponctuellement (copies vers scratch)**
  - IO/s mesurés aussi : < 125k : non saturants pour Myria



Problèmes rencontrés / Remarques

- Obligation d'arrêter les calculs avec des joblines monopolisant un grand nombre de nœuds (10 ici) avec simplement que quelques CPUs d'opérationnels (terminaison de collections).
  - Différences entre les collections : nombre de composés et difficulté de traitement des composés ?
- Run autour de 350 millions de composés max (difficile d'aller au dessus)
  - Vu le nombre de fichiers , il fallait créer des archives (CINES).
    - Très difficile de créer une archive (.tar) avant le transfert vers le CRIANN (CINES).
      - Espace disque associé qui est très grand et temps pour la réaliser.
    - Utilisation de la commande scp pour copier les fichiers.
      - Difficile d'être sûr à 100% de la qualité du transfert (possibilité de pertes (très rare cependant mais je l'ai observé)).
        - Accord entre le dépôt au CRIANN et les données au CINES avant suppression.

Protocole en lien avec Jedi Challenge

- Protocole avec VF.
  - Criblage primaire sur l'ensemble des structures.
    - Exhaustiveness = 1.
    - Sélection de près de 1 million de structures.
  - Criblage secondaire sur les molécules sélectionnées.
    - Exhaustiveness = 8 / flexibilité des résidus polaires sur le site de fixation.
    - Sélection de près de 50000 structures.

Résultats

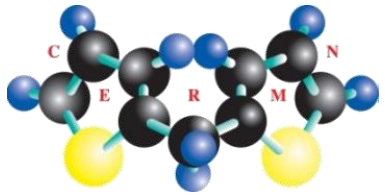
Criblage primaire (Exhaustiveness =1)								
3CLpro			**PLpro**			**TMPRSS2**		
559887710 composés dockés (90%)			567746375 composés dockés (91,5%)			586252945 composés dockés (94,5%)		
Min	Max	Median	Min	Max	Median	Min	Max	Median
-11.5	400.3	-7.1	-13.6	355.7	-4.8	-11.8	127.5	-6.7
930619 composés (cutoff à -9)			1106552 composés (cutoff à -10.1)			1185453 composés (cutoff à -9.0)		
Criblage secondaire (Exhaustiveness = 8 + flexibilité de certains résidus)								
57038 composés (cutoff à -9.5)			50616 composés (cutoff à -10.7)			58407 composés (cutoff à -9.6)		
Clustering								
10613 composés			12227 composés			9009 composés		

CINES : 2316227 heures de calcul CPUs
CRIANN : 1234858 heures de calcul CPUs
Autour de 8s par composé

CRIANN :
Volume : 120 To (30 juin) ; 9,6 To (31 juillet)
Nombre de fichiers : 42,6 millions (30 juin) ; 7,3 millions (31 juillet)

Conclusion

- L’analyse de milliards de composés par des techniques de docking semble être une possibilité de plus en plus intégrée comme approche pour trouver de nouveaux ligands par des méthodes *in silico*.
  - Avantage :
    - Une exploration forte de l’espace chimique avec des techniques complexes.
      - Cependant un niveau de simplification à intégrer pour le criblage primaire mais avec la sélection d’un sous-ensemble conséquent (autour du million de composés).
- Les alternatives (ou les approches précédentes) : on fait l’inverse.
  - Prétraitement des structures par des techniques chemoinformatiques.
  - Classer les molécules en groupes et ne tester que les centroïdes par docking et quelques représentants.
    - Tester les familles par docking qui répondent le mieux.
- Tendance : tester le plus grand nombre de composés par des méthodes complexes en s’aidant de l’augmentation de la puissance de calcul actuelle.



UNIVERSITÉ CAEN NORMANDIE

RÉGION NORMANDIE

Remerciements

CRIANN

GENCI
Le calcul intensif au service de la connaissance

INES
Centre Informatique National de l'Enseignement Supérieur