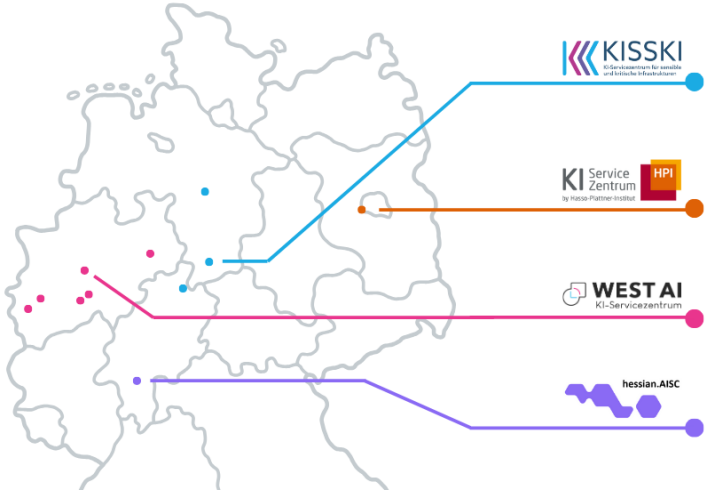
Hauke Kirchner

Generative KI-Dienste der GWDG

DLR

Inhalt

Eines von 4 nationalen BMBF geförderten KI Servicezentren



Uni Göttingen | GWDG | CIDAS | UMG | aQua | LUH | L3S | LiFE | LUIS | MHH | Fraunhofer IEE
Sensible und kritische Infrastrukturen
Medizin & Energie

Hasso-Plattner-Institut
Bildungs- und Beratungsangebote,
Einsatz von KI in Wirtschaft &
Gesellschaft

Uni Bonn | Fraunhofer IZB | Forschungszentrum Jülich | RWTH Aachen | TU Dortmund | Uni Paderborn
KI-Hardware, KI-Beratung, KI-Schulungen sowie multimodale und transferierbare KI-Modelle

TU Darmstadt
Erklärbarkeit, Generalisierbarkeit und kontextuelle Anpassung

KI-Servicezentrum KISSKI

KISSKI: KI Servicezentrum für Sensiblen und Kritischen Infrastrukturen

KISSKI
KI-Servicezentrum für sensible und kritische Infrastrukturen

https://kisski.de

- Erfüllung **Anforderungen** kritischer Infrastrukturen: Sicherheit, Datenschutz, Zuverlässigkeit
- gefördert mit 23 Mio € - bis 12/2027
  - KI-Dienste über GWDG zertifiziert gemäß **ISO 27001** (Informationssicherheit) - **C5 Testat** in Arbeit
  - Dienstleistungen für Pilotprojekte in ganz Deutschland
    - Beratung, Training, von Infrastruktur bis Endnutzerdienste
  - **Forschung** zur weiteren Verbesserung der Dienstleistungen
  - Fokus auf die Branchen Medizin und Energiewirtschaft ABER branchenoffen

Niederschwelliger Zugang - Quick-Start Guide für Neukund:innen

- Einstiegsberatung
- Monatliche KISSKI Online-Einführung
- Direkte Buchung via Webseite
  1. Registrierung in der Academic Cloud
  2. Auswahl der gewünschten Leistung im Leistungskatalog
  3. Buchung erfolgt auf der Leistungsseite
     - Angabe spezifischer Vorab-Informationen
- Monatliche Community Treffen bspw.
  - GöAID - Community für AI Developers
- Handling von Supportanfragen via Tickets

KISSKI Leistungsspektrum

- Ganzheitliches Angebot rund um KI
- KI Lösungen im Browser und via API
- Hardwareressourcen
  - Trainingsplattform (Nvidia H100, A100)
  - Future Technology Plattform (Intel Habana Gaudi 2, NVIDIA Grace Hopper, GraphCore, Esperanto.ai, SpiNNaker)
- Schulungsangebot HPC und KI
  - GWDG Academy
  - KISSKI Schulungsangebot
  - Individuelle Kurse auf Anfrage

hauke.kirchner@gwdg.de

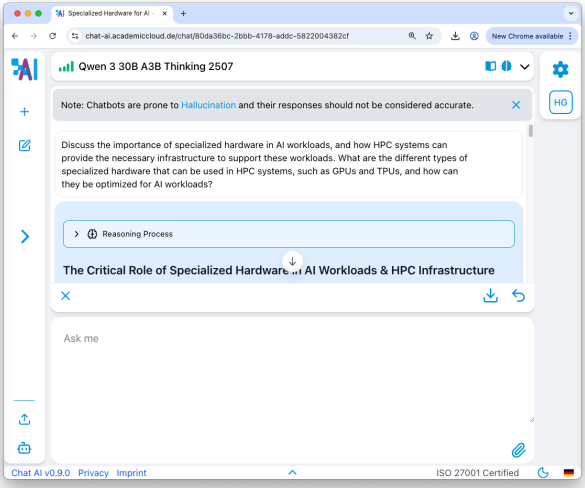
Generative KI-Dienste für Endanwender und gehostete Dienste

- KI Lösungen nutzbar im Browser
  - Chat AI, RAG, Voice AI, Image AI
  - API Zugang zu KI Modellen

- chat-ai.academiccloud.de
- voice-ai.academiccloud.de
- image-ai.academiccloud.de
- protein-ai.academiccloud.de
- docs.hpc.gwdg.de/services/coco

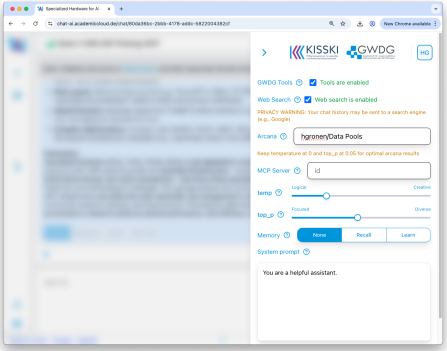
Chat AI

- Zugang zu derzeit 14 Open Source LLMs (und ChatGPT)
- **Konfigurierbar**: Systemprompt, `temp`, `top_p`
- **Import & Export** von Konversationen
- Einfaches Sharing von guten Einstellungen & Prompts
  - teilbar als "Personas"
- Anzeige der Modelverfügbarkeit



Was kann Chat AI ?

- Textverarbeitung in mehreren Sprachen inklusive Verarbeitung von PDF & json
- Bild- und Videoverarbeitung (Vision Models)
- Sprachverarbeitung
- Reasoning, z.B. DeepSeek
  Ausgabe des "Denkprozesses" des Models
- Retrieval Augmented Generation (RAG)
  Nutzung eigener Dokumentensammlung als Grundlage

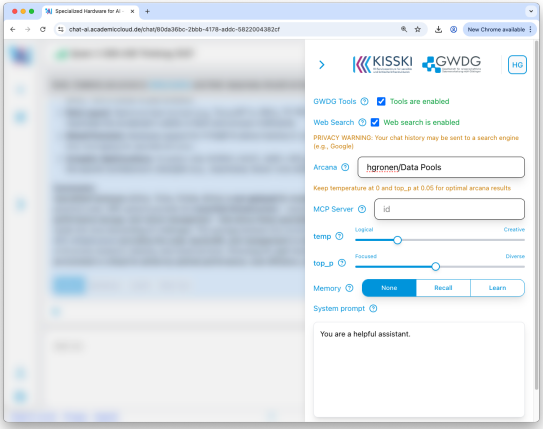


Modelkatalog

- Stetig aktualisierter Modelkatalog
- Spezialisierte Modelle (Stand: 5.9.2025)
  - **Schnellstes** - Llama 3.1 8B Instruct
  - **Vision Model** - Gemma 3 27B Instruct, InternVL2.5 8B MPO, MedGemma 27B Instruct, Qwen 2.5 VL 72B Instruct
  - **Multilingual** - Qwen 3 32B, Qwen 2.5 VL 72B Instruct, Mistral Large Instruct
  - **Deutsch** - Llama 3.1 SauerkrautLM 70B Instruct
  - **Reasoning** - Qwen 3 235B A22B Thinking 2507, Qwen QwQ 32B, DeepSeek R1 0528, DeepSeek R1 Distill Llama 70B, Llama 3.3 70B Instruct, Mistral Large Instruct
  - **Coding** - Mistral Large Instruct, Codestral 22B, Qwen 2.5 Coder 32B Instruct
  - **RAG** - Qwen 3 30B A3B Thinking 2507, Llama 3.1 8B RAG, Llama 3.1 SauerkrautLM 70B RAG

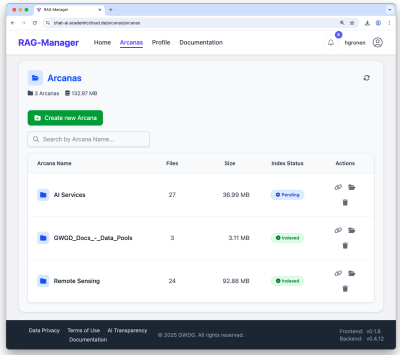
Retrival Augmented Generation (RAG)

- Integriert in Chat AI
  - Server-Seitig auf GPU-Knoten
- Vorgehen des RAG-Systems
  - Informationsextraktion aus benutzerdefinierten Dokumenten
  - LLM kann darauf basierend eine Antwort generieren



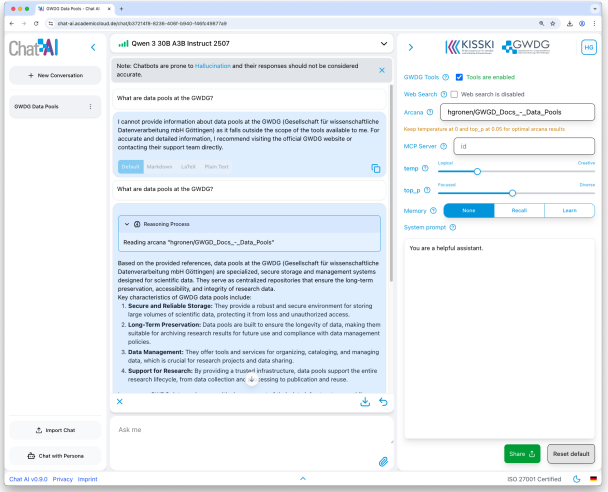
RAG Management Interface

- Erstellung von Dokumenten-sammlungen, "Arcanas"
- Unterstützung von Text, Markdown und PDF Dateien
- Hinzufügen neuer Daten mittels Embedding Prozess
- Teilen von Arcanas mittels Link



RAG Integration: Chat Interface

- Chat AI nutzt unsere RAG-Middleware
- Lädt bei Anfrage Daten aus Dokumentensammlungen
- Prompt wird mit Ausschnitten aus Dokumenten erweitert
- Chat AI erzeugt Antwort
- Funktioniert exklusiv mit internen Modellen

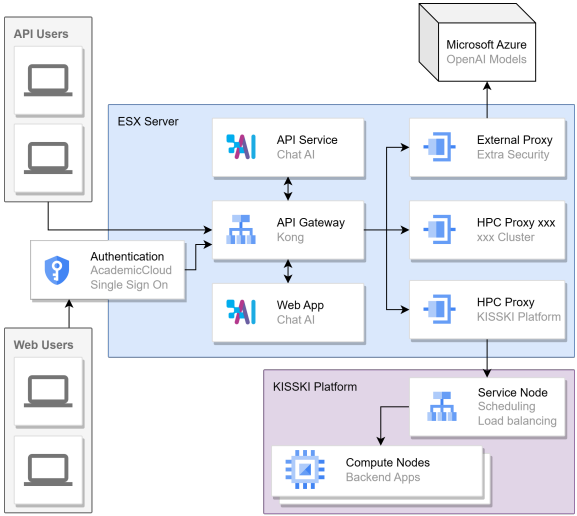


Weiterentwicklung: Scalable AI Accelerator
SAIA Plattform und Ökosystem

- Stand: Die bisherigen Lösungen für Gen-KI-Dienste wachsen zusammen
- Ziel: Entwicklung zur **Plattform** für KI Inferenz Services
  - Für eigene Services: Chat AI, CoCo AI, Voice AI, Image AI, ...
  - Nutzende können auch eigene Services hosten
  - Können über KISSKI oder an anderen HPC Zentren gehostet werden
  - SDK für Entwickung an beliebigen Zentren / lokal
- Zusätzliche **Beratung** zur Bereitstellung und Anwendung von KI
  - Benutzerdefinierte Daten in LLM via RAG
  - Strategien zur Einhaltung des Datenschutzes (EU AI Act)
- **Informationssicherheit**: ISO 27001 zertifizierte Plattform
- Hosting der Modelle von Trainingsplattform und mehr!

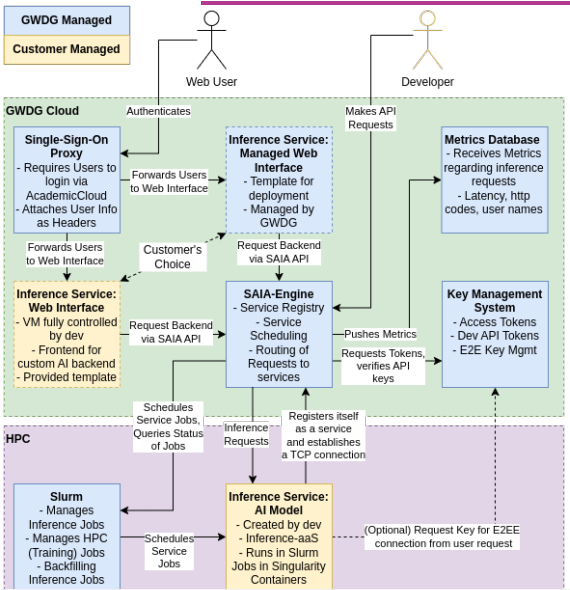
SAIA 0.9 - Aktuelle Architektur - Interne Abläufe

- Proxy übernimmt Authentifizierung
- API Gateway wickelt gesamte interne Routing ab
- HPC Proxy stellt SSH-Tunnel zu HPC
- Verwendet SSH ForceCommand, fungiert als "circuit breaker"
- Bspw. vLLM läuft in Slurm-Jobs



SAIA 2.0 Architecture

- Kundenservice mit Frontend in Cloud und Backend auf HPC
- Web interface auf Cloud VM
- KI Modelle von Kunden in Singularity Containern auf HPC
- Anfragen Ende-zu-Ende Verschlüsselt über KMS
- SAIA als Broker nutzen und Integrität garantieren



SAIA 2.0

- Weiterentwicklung der Lösung in 2025
- **Ziel**: Platform-as-a-Service (PaaS) für Inferenz
- Bereitstellung von KI services auf unserer SAIA Plattform durch Kunden
  - Benutzerfreundliches Hosting von Inferenz-Modelle
  - Zusätzlich: Hosting von Webinterfaces/Frontends für Inferenz-Modelle
- Vorteile:
  - Self-Service
  - Automatische Skalierung
  - Automatisches Monitoring
  - Nutzung von HPC-Hardware
  - Nutzung von Accounting und User-Management
- API Key self-service, Developer Interface: `https://saia.gwdg.de/`

Benutzerdefinierte Anwendungen

- OpenAI-kompatibler API Schlüssel kann erstellt werden
- Kann in viele Anwendungen integriert werden, zB.
  - AnythingLLM, RAG auf dem Desktop
  - LlamaIndex, RAG Bibliothek
  - DSPY, LLM Programming Frameworking
  - SillyTavern, Role-play chatting
  - https://github.com/Hannibal046/Awesome-LLM
- Self hosting von eigenen ChatBots, Chat Frontends, zB.
  - Bibliothek von Dokumenten für Studierende
  - FAQs als ChatBot
  - Automatisierter Support auf Grundlage alte Antworten
- Templates für Modellausführung auf HPC für eigene Forschung

Rechtliche Nutzungsvoraussetzung

- **Rahmenvertrag** mit GWDG möglich
  - Deckt bis Schutzstufe C ab
  - Derzeit Anpassung auf Schutzstufe D (ISO 27701)
- Verarbeitung von Patientendaten über **Secure HPC** Workflow
- Chat AI Services aktuell noch unter Schutzstufe C
  Keine Eingabe sensibler Daten
- Erwerb **C5-Testat** geplant

Nutzungsverträge für AI-Service

Chat AI
interne (Open-Weight) Modelle

- Ausgefüllte und unterschriebene **Leistungsvereinbarung**
  Academic Cloud Basis
  + Dienstleistungspaket Chat AI
- ausgefüllter und unterschriebener **Auftragsverarbeitungsvertrag**
- Ausgefülltes Formular zur **Benennung des Datenschutzbeauftragten**
- ausgefülltes und unterschriebenes Formular zur **Ernennung von IDM-Fachverantwortlichen**
- Optional: **Chat AI +**
  externe (kommerzielle) Modelle
- API-Keys zur Integration in eigene Frontends auf Anfrage

Open AI Modelle

- GPT-5 Chat
- GPT-5 Mini
- GPT-5 Nano
- o3
- o3 mini
- more in our documentation

- **Pauschalbetrag**
  Kosten per Token entsprechen Azure Region Sweden Central (ab 250 € netto monatlich)
- **Deckelbetrag**
  Einschränkung des Zugangs nach Erreichen eines Deckelbetrags
- **Tatsächliche Nutzung**
  Berechnung Tokens monatsaktuell
- **Mehraufwand**
  Zusätzliche Rechnung wenn Pauschalbetrag überschritten wid

Image AI

- Text-to-image Generierung
- Erstes Model `FLUX.1-schnell`
- Basiert auf OpenAI-kompatiblem API Server
- Optional: Image-to-image
- `https://image-ai.academiccloud.de/`

prompt: "A high performance computing cluster. In the Background a cat sitting on top of the HPC cluster and holding a sign that says ‘Coming soon!’. At the top ‘GWDG’, in clean, simple, light blue letters. In the center of the image ‘Image AI’ in clean, simple, light blue letters."

Image AI

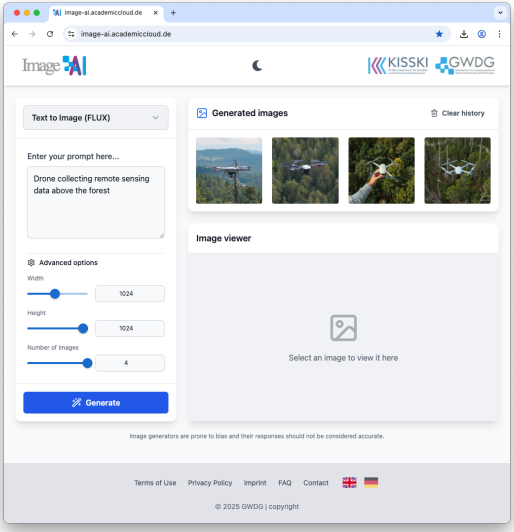


image-ai.academiccloud.de
Image AI
KISSKI
GWDG
Text to Image (FLUX)
Enter your prompt here...
Drone collecting remote sensing data above the forest
Advanced options
Width
1024
Height
1024
Number of Images
4
Generate
Generated images
Clear history
Image viewer
Select an image to view it here
Image generators are prone to bias and their responses should not be considered accurate.
Terms of Use
Privacy Policy
Imprint
FAQ
Contact
© 2025 GWDG | copyright

Voice AI

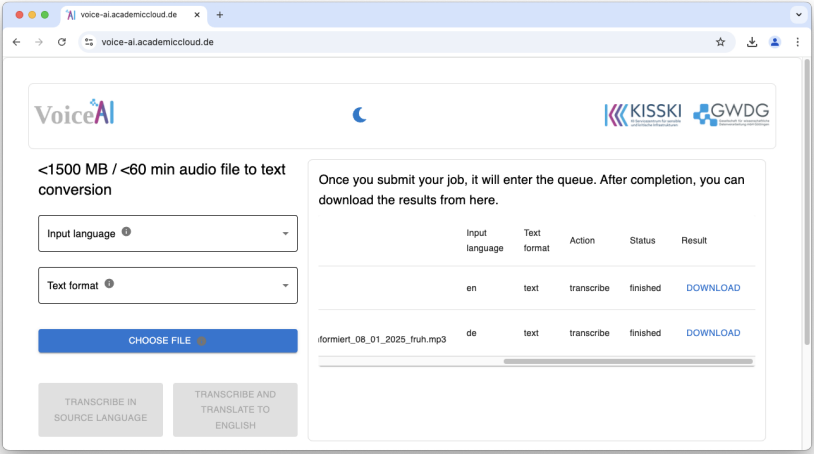
Transkription & Translation

- Transkribiert Audiodateien in Text
- Erzeugt Videountertitel
- Unterstützt mehrere Eingabesprachen
  - Transkribiert in Eingabesprache
  - Übersetzt ins Englische
- Kann über API genutzt werden

BBB Integration Prototyp

- Besprechungen in Echtzeit transkribieren
- Nahtlose Integration in BBB
- Sitzungszusammenfassung
- Erhöht die Inklusivität Barrierefreiheit

Voice AI


voice-ai.academiccloud.de
VoiceAI
KISSKI
GWDG
<1500 MB / <60 min audio file to text conversion
Input language
Text format
CHOOSE FILE
TRANSCRIBE IN SOURCE LANGUAGE
TRANSCRIBE AND TRANSLATE TO ENGLISH
Once you submit your job, it will enter the queue. After completion, you can download the results from here.
Input language
Text format
Action
Status
Result
en
text
transcribe
finished
DOWNLOAD
iformiert_08_01_2025_fruh.mp3
de
text
transcribe
finished
DOWNLOAD

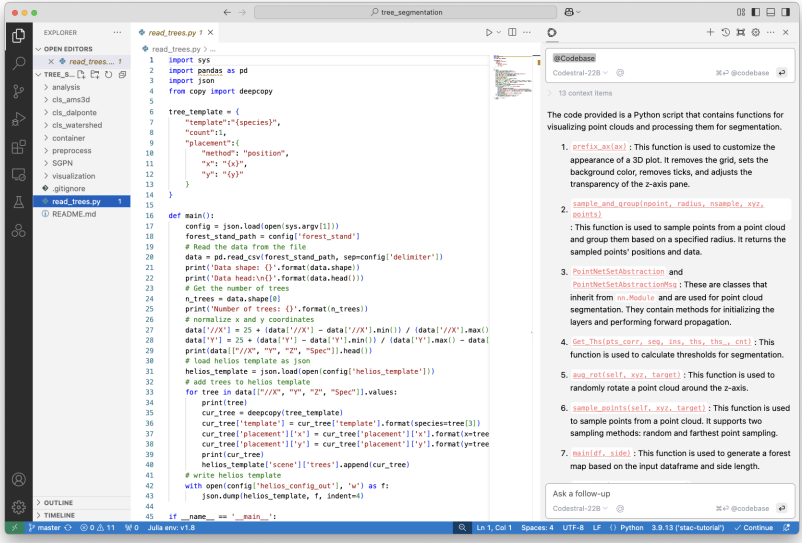
CoCo AI

- Code-Vervollständigungsdienst
- Hilft bei Bearbeitung, Erzeugung, Korrektur und Kommentierung von Code
- Integrierbar in **VS Code** und **JetBrains** mittels des Continue-Plugin
- Zugriff auf alle Chat AI LLMs
- Gleiche Sicherheit und Datenschutz wie Chat AI
- `https://docs.hpc.gwdg.de/services/coco/index.html`

Was kann CoCo AI ?

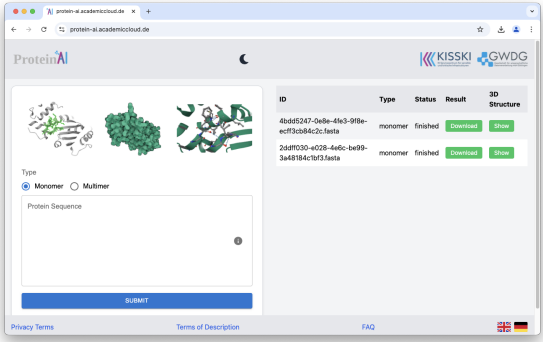
- **Analysieren**
  - Verwendet Snippet / Datei / Codebase als Kontext
- **Generieren**
  - Generierung von Code basierend auf Kontext und Aufgabe
- **Reparieren**
  - Schlägt Code oder Befehle zur Fehlerbehebung vor
- **Vervollständigen**
  - Schlägt Code zur Ergänzung der aktuellen Zeile vor

CoCo AI

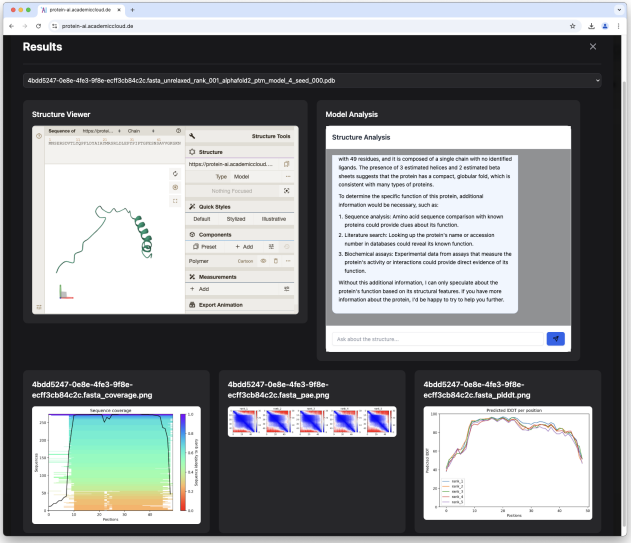

tree_segmentation
EXPLORER
OPEN EDITORS
read_trees.py 1
TREE_S...
analysis
cls_ams3d
cls_dalponte
cls_watershed
container
preprocess
SGPN
visualization
.gitignore
read_trees.py 1
README.md
OUTLINE
TIMELINE
read_trees.py 1 ×
read_trees.py > ...
import sys
import pandas as pd
import json
from copy import deepcopy
tree_template = {
"template":"{species}",
"count":1,
"placement":{
"method": "position",
"x": "{x}",
"y": "{y}"
}
}
def main():
config = json.load(open(sys.argv[1]))
forest_stand_path = config['forest_stand']
# Read the data from the file
data = pd.read_csv(forest_stand_path, sep=config['delimiter'])
print('Data shape: {}'.format(data.shape))
print('Data head:\n{}'.format(data.head()))
# Get the number of trees
n_trees = data.shape[0]
print('Number of trees: {}'.format(n_trees))
# normalize x and y coordinates
data['//X'] = 25 + (data['//X'] - data['//X'].min()) / (data['//X'].max()
data['Y'] = 25 + (data['Y'] - data['Y'].min()) / (data['Y'].max() - data[
print(data[["//X", "Y", "Z", "Spec"]].head())
# load helios template as json
helios_template = json.load(open(config['helios_template']))
# add trees to helios template
for tree in data[["//X", "Y", "Z", "Spec"]].values:
print(tree)
cur_tree = deepcopy(tree_template)
cur_tree['template'] = cur_tree['template'].format(species=tree[3])
cur_tree['placement']['x'] = cur_tree['placement']['x'].format(x=tree
cur_tree['placement']['y'] = cur_tree['placement']['y'].format(y=tree
print(cur_tree)
helios_template['scene']['trees'].append(cur_tree)
write helios template
with open(config['helios_config_out'], 'w') as f:
json.dump(helios_template, f, indent=4)
if __name__ == '__main__':
@Codebase
Codestral-22B
@codebase
13 context items
The code provided is a Python script that contains functions for visualizing point clouds and processing them for segmentation.
1. prefix_ax(ax) : This function is used to customize the appearance of a 3D plot. It removes the grid, sets the background color, removes ticks, and adjusts the transparency of the z-axis pane.
2. sample_and_group(npoint, radius, nsample, xyz, points) : This function is used to sample points from a point cloud and group them based on a specified radius. It returns the sampled points' positions and data.
3. PointNetSetAbstraction and PointNetSetAbstractionMsg : These are classes that inherit from nn.Module and are used for point cloud segmentation. They contain methods for initializing the layers and performing forward propagation.
4. Get_Ths(pts_corr, seg, ins, ths, ths_, cnt) : This function is used to calculate thresholds for segmentation.
5. aug_rot(self, xyz, target) : This function is used to randomly rotate a point cloud around the z-axis.
6. sample_points(self, xyz, target) : This function is used to sample points from a point cloud. It supports two sampling methods: random and farthest point sampling.
7. main(df, side) : This function is used to generate a forest map based on the input dataframe and side length.
Ask a follow-up
Codestral-22B
@codebase
master
Julia env: v1.8
Ln 1, Col 1
Spaces: 4
UTF-8
LF
Python
3.9.13 ('stac-tutorial')
Continue

Protein AI

- Ziel: Vorhersage von 3D Strukturen der Proteine
- **Code:** Colabfold
  - Umfangreiche Datenbanken
  - Schnelles und empfindliches MMseqs2-Verfahren
  - Vergleichbare Ergebnisse zu Alphafold2



Protein AI


Results
4bdd5247-0e8e-4fe3-9f8e-ecff3cb84c2c.fasta_unrelaxed_rank_001_alphafold2_ptm_model_4_seed_000.pdb
Structure Viewer
Structure Tools
Structure
Quick Styles
Components
Measurements
Export Animation
Model Analysis
Structure Analysis
with 49 residues, and it is composed of a single chain with no identified ligands. The presence of 3 estimated helices and 2 estimated beta sheets suggests that the protein has a compact, globular fold, which is consistent with many types of proteins.
To determine the specific function of this protein, additional information would be necessary, such as:
1. Sequence analysis: Amino acid sequence comparison with known proteins could provide clues about its function.
2. Literature search: Looking up the protein's name or accession number in databases could reveal its known function.
3. Biochemical assays: Experimental data from assays that measure the protein's activity or interactions could provide direct evidence of its function.
Without this additional information, I can only speculate about the protein's function based on its structural features. If you have more information about the protein, I'd be happy to try to help you further.
Ask about the structure...
4bdd5247-0e8e-4fe3-9f8e-ecff3cb84c2c.fasta_coverage.png
4bdd5247-0e8e-4fe3-9f8e-ecff3cb84c2c.fasta_pae.png
4bdd5247-0e8e-4fe3-9f8e-ecff3cb84c2c.fasta_plddt.png

Community

- GöAID, Monatliches Treffen für AI
- AI Community (Matrix Channel)
- Chat AI Support (Matrix Channel)
- Chat AI Mailing Liste
- SAIA Mailing Liste