

# **Analysing Granted Projects of the European Research Council (ERC) to Understand Where Science Money Goes**

Leor Ariel Rose and Aviv Rovshitz

Department of Software and Information Systems Engineering

Ben-Gurion University of the Negev Beer-Sheva, Israel

{leorro, avivrov}@post.bgu.ac.il

## **1. Introduction**

Modern research progress relies heavily on research funding allocation. Developing a deep understanding of how funding is distributed and how it impacts scientific research is of the utmost importance. This is to ensure accountability and foster transparency within the scientific community. In this paper, we examine the expenditure of funds on research projects funded by the prestigious European Research Council (ERC). In order to uncover patterns, trends, biases, and impacts of funding allocations, we combine various external information sources and use a variety of analyses such as statistics, network analysis, geographic analysis, and natural language processing. Our results show an extensive analysis and discover new information on the grant awareness in ERC.

## **2. Results**

### **2.1. Dataset**

We use a public dataset from ERC on projects funded by the ERC. The dataset contains the following features: Project number, Project acronym, Project title, Abstract, PI, Grant type, Topic, Project budget, Call ID, Host Institution, City, Country and 14474 records.

### **2.2. Univariate statistical analysis**

Univariate analysis revealed several interesting insights regarding the types of grants, funding topics, institutions, and countries that dominate the research landscape. The key findings are summarised as follows: **(1)** majority of grants fall under the category of "Starting Grants" (Supplementary Information, Figure 1), **(2)** significant funding is directed to four topics - Products & Processes Engineering, Synthetic Chemistry & Materials, Prevention, Diagnosis & Treatment of Human Diseases, and Computer Science & Informatics (Supplementary Information, Figure 2), **(3)** the National Centre for Scientific Research (CNRS) stands out as the foremost recipient of funded grants (Supplementary Information, Figure 3), **(4)** the UK, Germany, and France led in funding (Supplementary Information, Figure 4), **(5)** grants with more years get more budget (Supplementary Information, Figure 5). For more multivariate analysis refer to the jupyter notebook attached.

### **2.3. Multivariate analysis**

Univariate analysis can reveal a lot of information but taking into consideration more than one feature we can learn connections and patterns in the data. The key findings of our multivariate analysis are summarised as follows: **(1)** distribution of grants across different types over the years revealed a consistent pattern of stability, with the exception of "Synergy grants" (Supplementary Information, Figure 6). **(2)** We observed a dynamic pattern in the annual budget changes received by host institutions. Notably, certain institutions, such as the National Centre for Scientific Research, Weizmann Institute of Science, and Swiss Federal Institute of Technology, consistently occupied positions within the top 30 funded host institutions (Supplementary Information, Figure 7). **(3)** Intriguing shifts in the budget distribution among recipient countries over time. For instance, the United Kingdom consistently ranked among the top 5 recipients from 2008 to 2021. However, a significant change occurred in 2022, as the UK's position dropped to the top 8. Strikingly, in 2023, the United Kingdom experienced a further decline and failed to secure a position within the top 30 recipients. Conversely, Germany, France, and the Netherlands demonstrated remarkable consistency, maintaining their positions within the top recipients every year throughout the observation period. For more multivariate analysis refer to the jupyter notebook attached (Supplementary Information, Figure 8).

### **2.4. Network analysis**

In the course of establishing a network of grants and conducting an in-depth analysis of the involved communities and cliques, several noteworthy findings emerged: **(1)** Among the grants observed, a range of collaborative arrangements was evident, involving multiple Principal Investigators (PIs) such as 2 PIs (32 grants), 3 PIs (61 grants), 4 PIs (30 grants), and 5 PIs (1 grant). Nonetheless, the predominant trend remained the allocation of grants to individual PIs (Supplementary Information, Figure 9). **(2)** The investigation unveiled a total of 133 cliques, including instances like 'University of Reading', 'Foundation for Research and Technology Hellas', 'University of Stuttgart', and 'University of Freiburg'. Although no particularly distinct cliques emerged, this still provided valuable insights into collaborative patterns within grant allocations. **(3)** The examination of communities further illuminated the presence of multiple distinct groupings. While these communities did not yield significant specific insights, they still shed light on the prevailing collaborative dynamics within the realm of grant distribution (Supplementary Information, Figure 10).

### **2.5. Natural language processing analysis**

When looking at abstract embeddings tied to money and topics (Supplementary Information, Figure 11), we found groups based on topics. But, the budget of these groups wasn't decided by the topic. Also, when we connected abstract embeddings to money and institutions (Supplementary Information, Figure 12), we didn't see clear groups based on institutions or budgets. This shows that the institution you're from doesn't decide how much budget you get

or how you write your abstract. We did the same with abstract embeddings, money, and countries (Supplementary Information, Figure 13). We didn't find clear groups based on countries or budgets. This means where you're from doesn't decide how much budget you get or how you write your abstract. We saw similar things with titles. Looking at money and topics, we found clear groups based on topics (Figure 1). But, the budget of these groups didn't change with the topic. Also, when we looked at money and institutions (Supplementary Information, Figure 14), we didn't see clear groups. This supports the idea that the institutions you're from doesn't really change how much budget you get or how you structure your title. Similarly, when we checked money and countries in titles (Supplementary Information, Figure 15), we didn't see clear groups based on countries or budgets. This shows that the country doesn't have a big say in how much budget you get or how you structure your title.

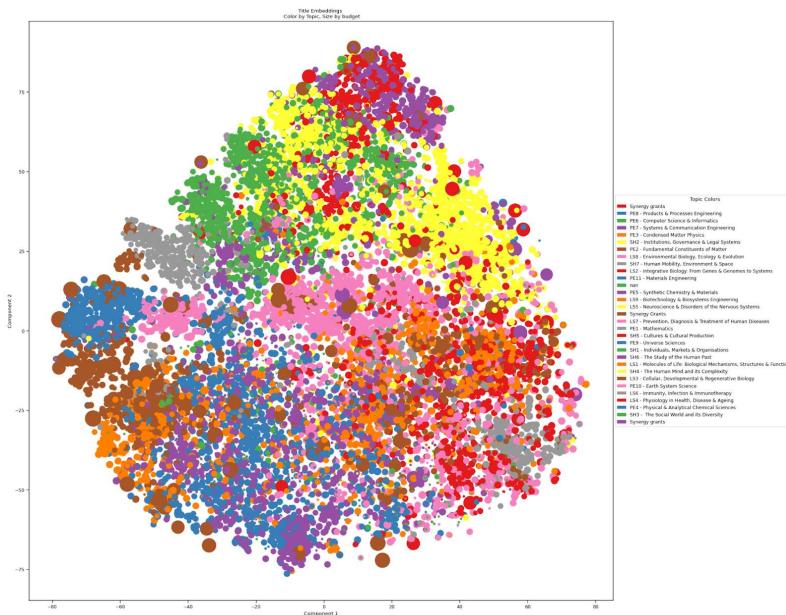


Figure 1: Grants title embeddings in relation to money (size) and topic (colour).

Additionally, an exploration of the most repeated words in grant titles (Supplementary Information, Figure 16) and abstracts (Supplementary Information, Figure 17) revealed common themes and terms used by researchers in their project descriptions. The most frequently repeated words were "dynamic," "system," "new," "novel," "cell," "human," "mechanism," and "molecular." These words highlight the emphasis on exploring dynamic systems, novel approaches, and cellular and molecular mechanisms in various research projects.

## 2.7. Grant amount prediction and explainability

We trained an XGBRegressor model on 80% of the data for predicting grant budget and evaluated on the remaining 20% (Supplementary Information, Figure 18). The mean squared error (MSE) on the test set was determined to be  $1.5676321\text{e}+11$ , indicating the model's

predictive performance. Additionally, a 95% confidence interval for the root mean squared error (RMSE) was calculated to be between 344858.34 and 439079.73, with a standard error of 25899.63 (Supplementary Information, Figure 19, 20). The feature importance ranking, from highest to lowest, is as follows: topic, host institution, call id, start year, grant type, country, city, and duration (Supplementary Information, Figure 21). These observations shed light on how individual features contribute to the model's predictions, enhancing our understanding of the underlying relationships within the data.

## **2.7. Integration of external information for a extended analysis**

In our experiment, we went beyond the basics by adding more information from different places. We looked at university ratings from Kaggle, data we collected from Google Scholar (like citations, publications, and coauthors), and even country currencies from a public GitHub file. After analysing everything, we found some interesting things. First, there's not much of a connection between how good a university is rated and the money they get from grants (Supplementary Information, Figure 23). The same goes for the currency of a country and grant money (Supplementary Information, Figure 24). When we looked at the number of papers a researcher had before they got a grant and the grant money they got, we saw some patterns. For "Proof of Concept" grants, more papers meant more money, while for "Synergy" grants, it was the opposite (Supplementary Information, Figure 25). Also, in "Proof of Concept" grants, more grant money meant more papers and citations (Supplementary Information, Figure 26, 27, 28). But the number of coauthors didn't really change with grant money (Supplementary Information, Figure 29). So, grant budgets and academics are kind of connected, but it's more complicated than we thought.

## **4. Conclusion**

In this work we presented an extensive analysis on the grants in ERC. We used a variety of techniques such as statistics, network analysis, geographic analysis, and natural language processing. Our results may not be the most significant or new but we show how an analysis on grant data can uncover patterns, trends, biases, and impacts of funding allocations. This work can open the door for future work combining more data and features in order to get a deep understanding of how funding is distributed and how it impacts scientific research.

## **5. Acknowledgment**

For text code assistance we used ChatGPT and Wordtune.

## **6. CRediT author statement**

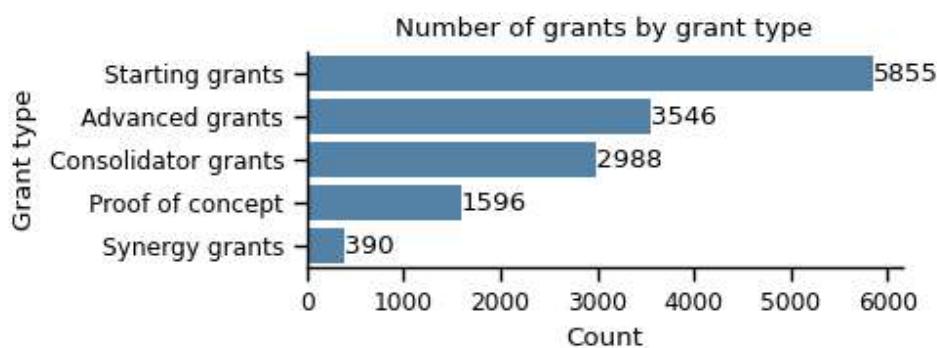
Leor Ariel Rose & Aviv Rovshitz : Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization, Project administration.

# Analysing Granted Projects of the European Research Council (ERC) to Understand Where Science Money Goes - Supplementary Information

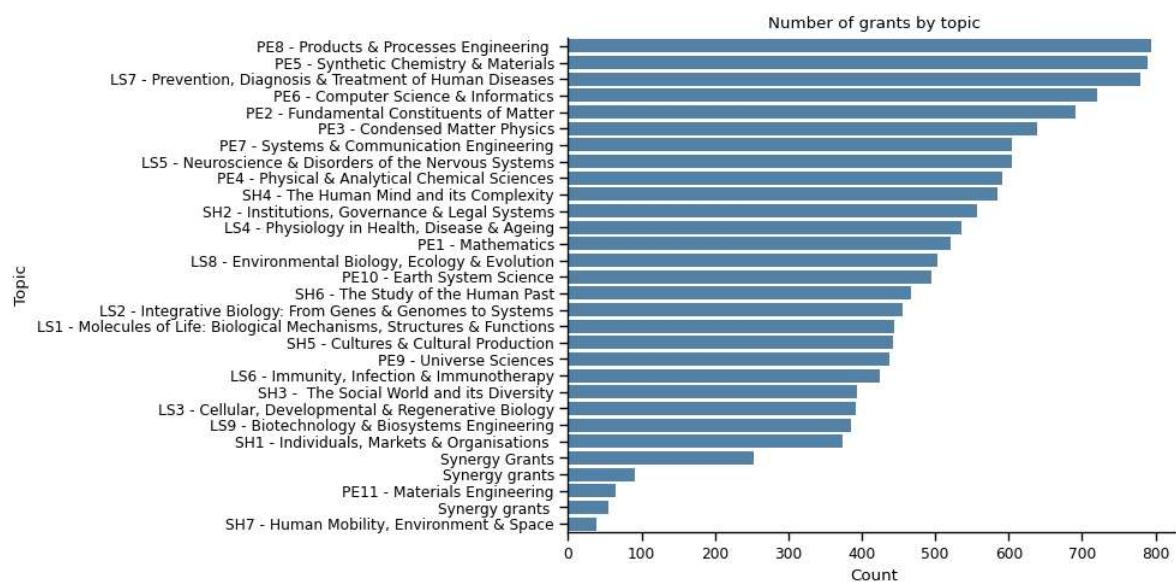
Leor Ariel Rose and Aviv Rovshitz

Department of Software and Information Systems Engineering  
Ben-Gurion University of the Negev Beer-Sheva, Israel  
[{leorro, avivrov}@post.bgu.ac.il](mailto:{leorro, avivrov}@post.bgu.ac.il)

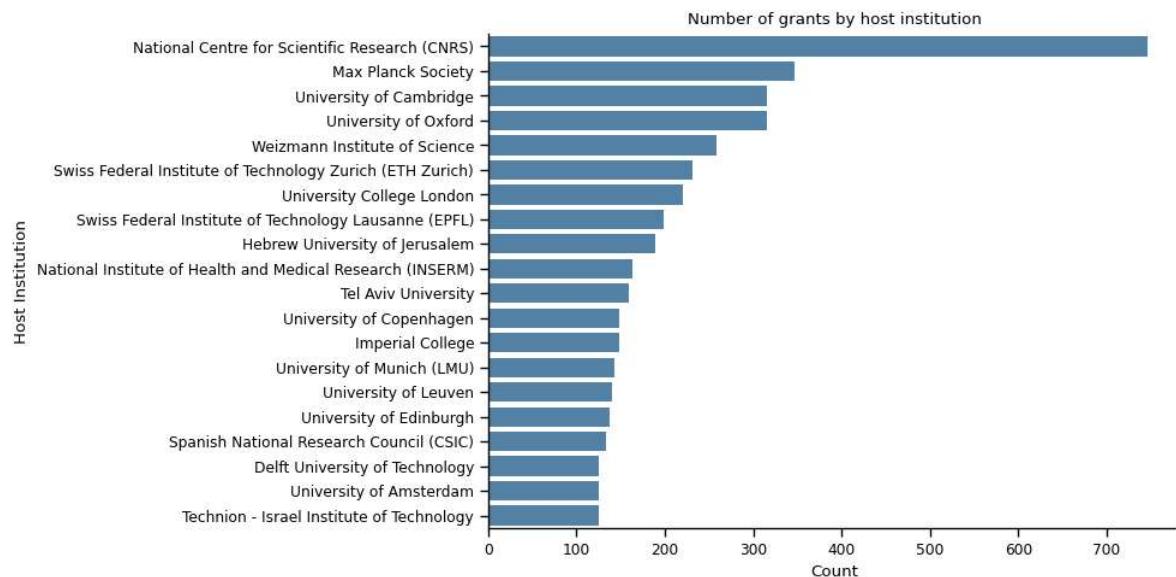
## Figures from univariate statistical analysis



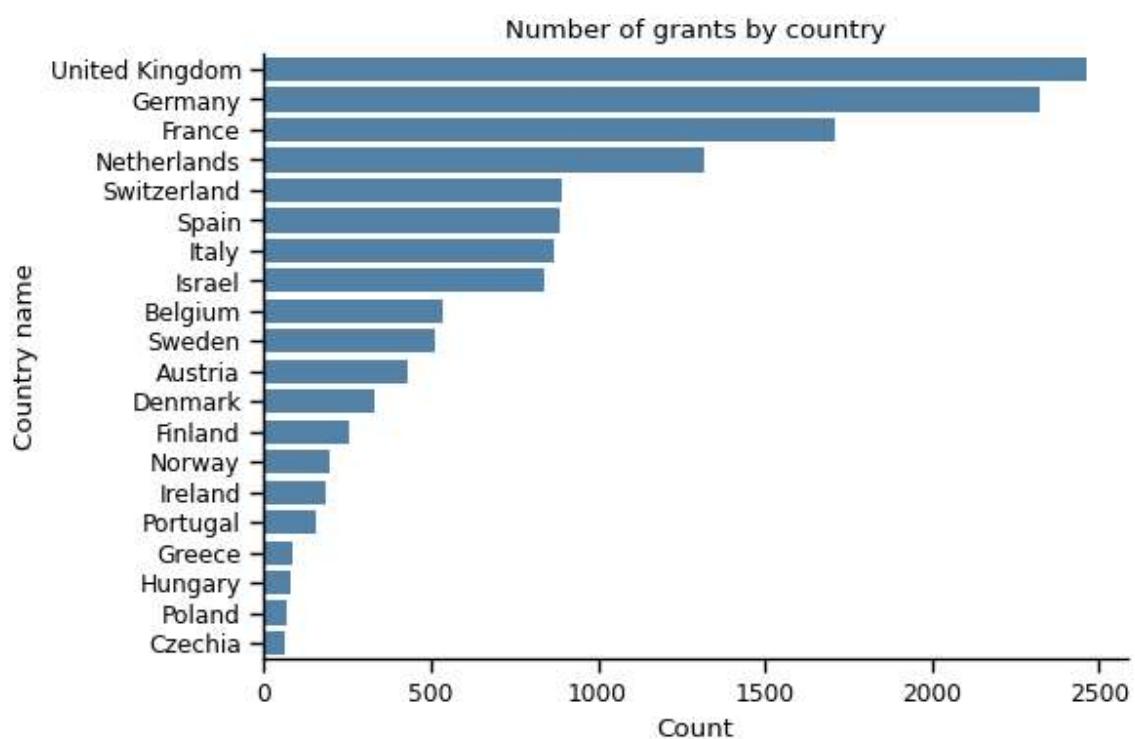
Supplementary figure 1: Distribution of grants per grant type for all grants in the ERC dataset.



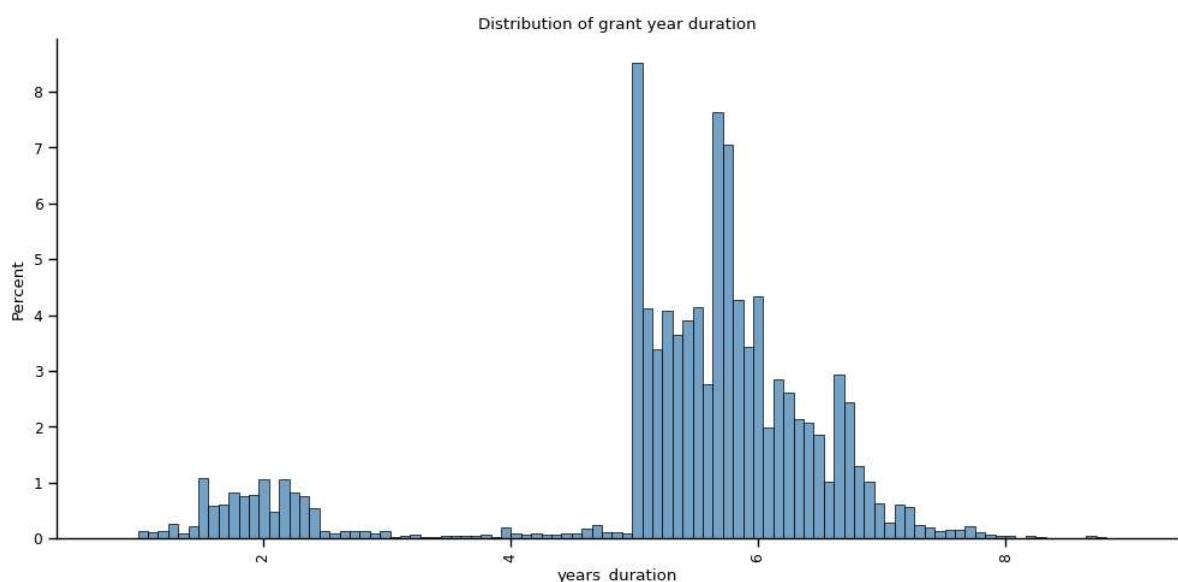
Supplementary figure 2: Distribution of grants per grant topic for all grants in the ERC dataset.



Supplementary figure 3: Distribution of grants per grant host institution for all grants in the ERC dataset.

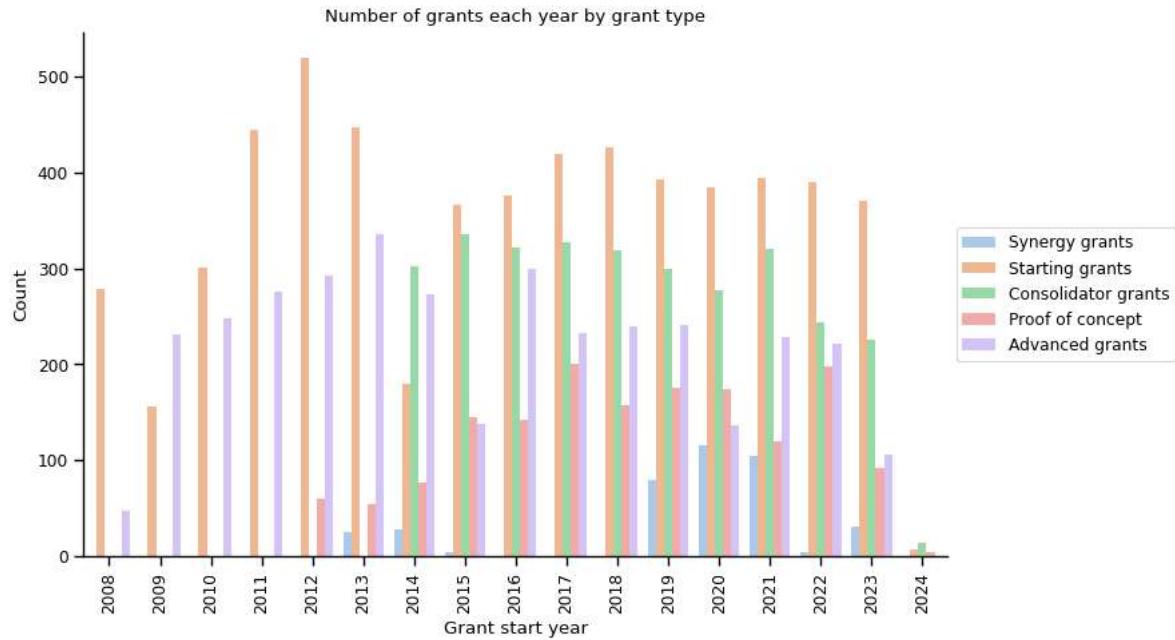


Supplementary figure 4: Distribution of grants per grant country for all grants in the ERC dataset.

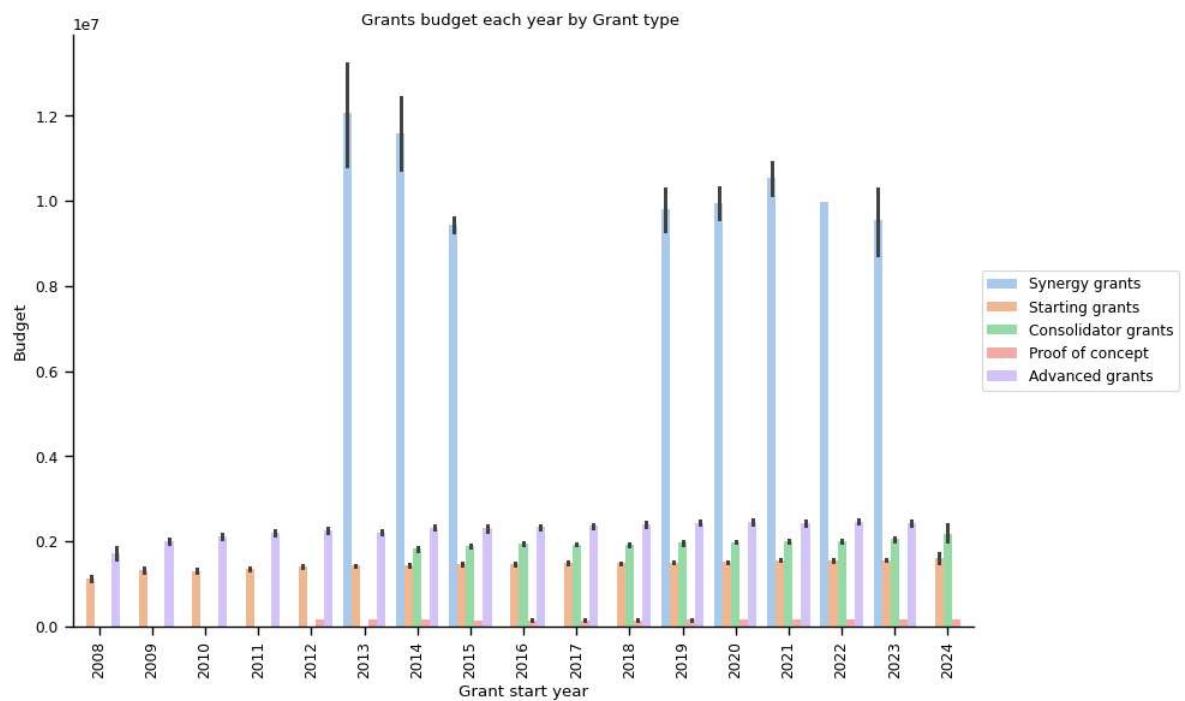


Supplementary figure 5: Distribution of grants per grant year duration for all grants in the ERC dataset.

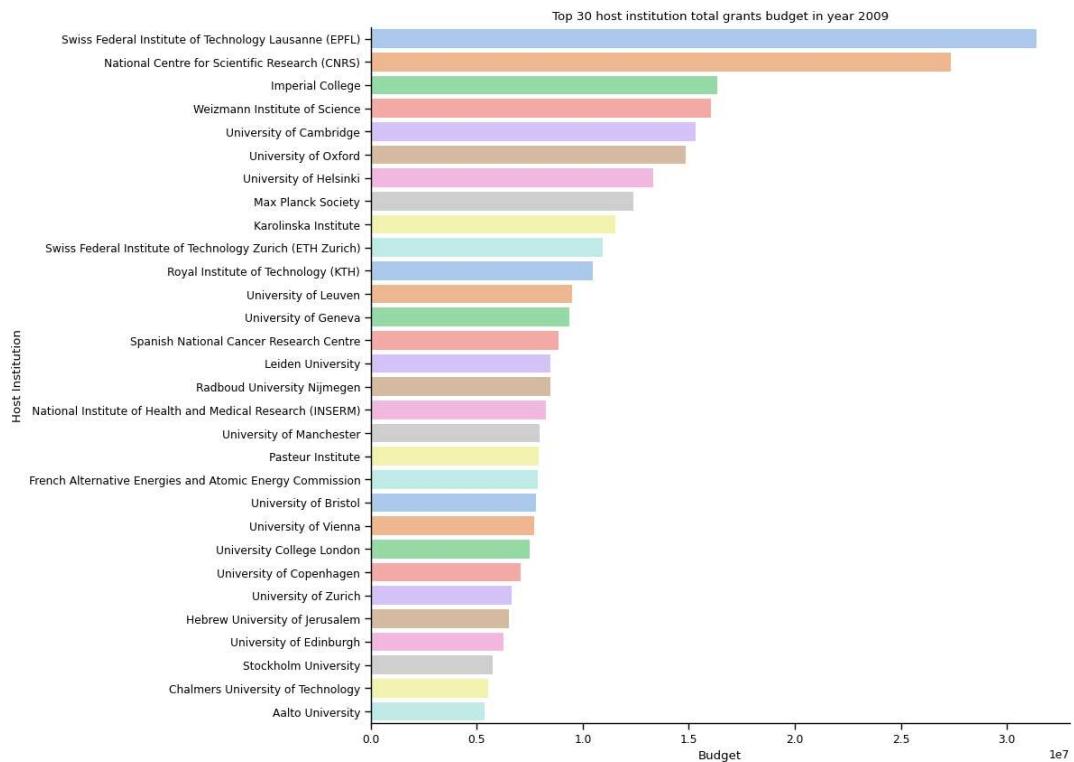
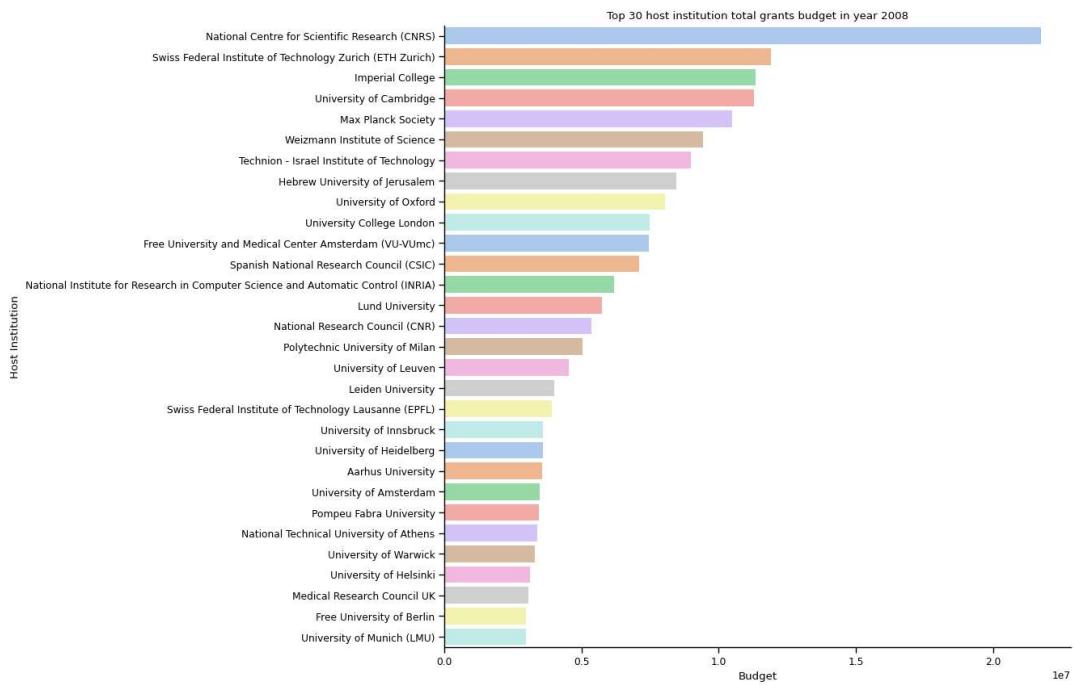
## Figures from Multivariate statistical analysis

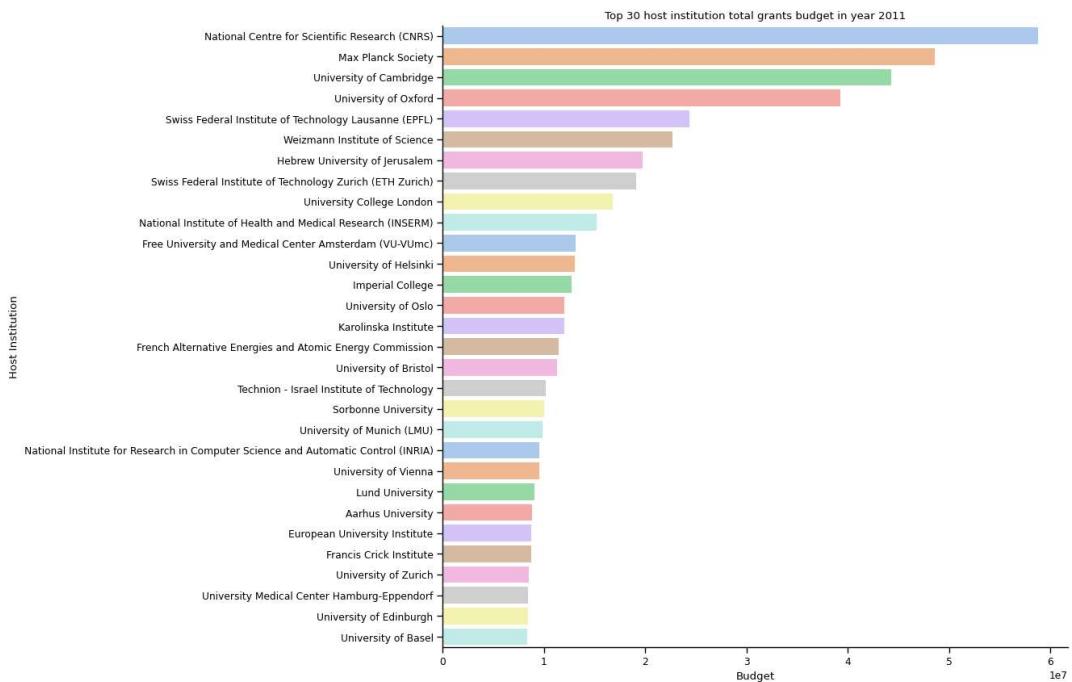
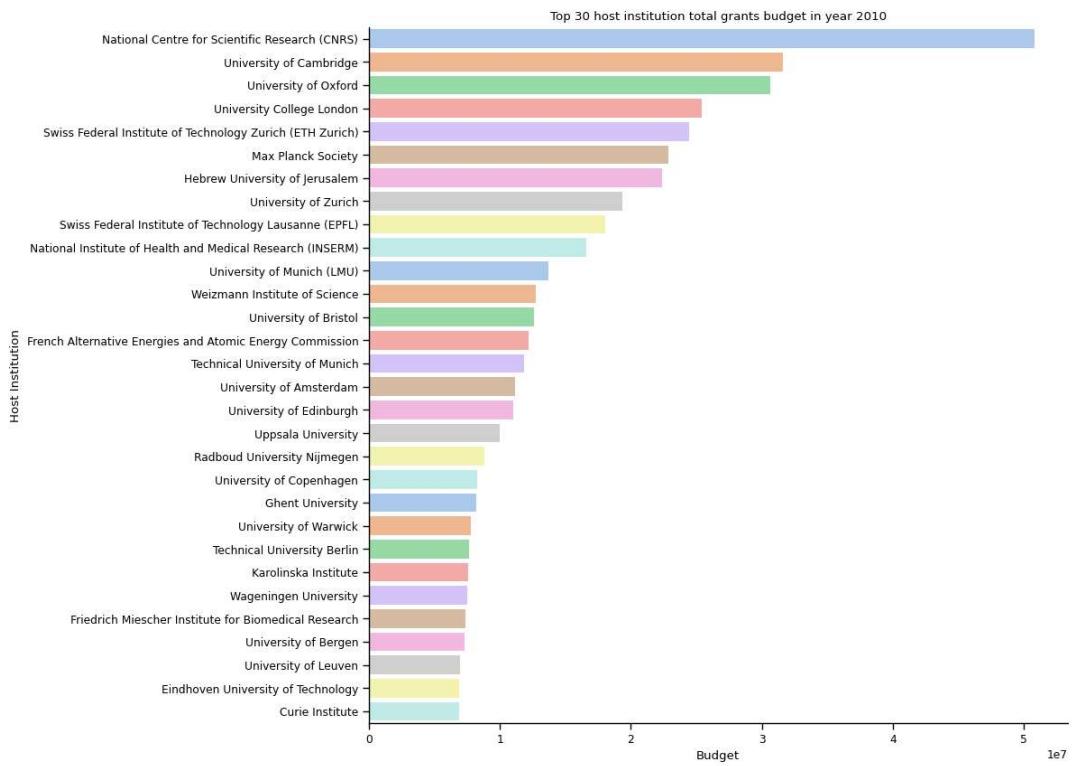


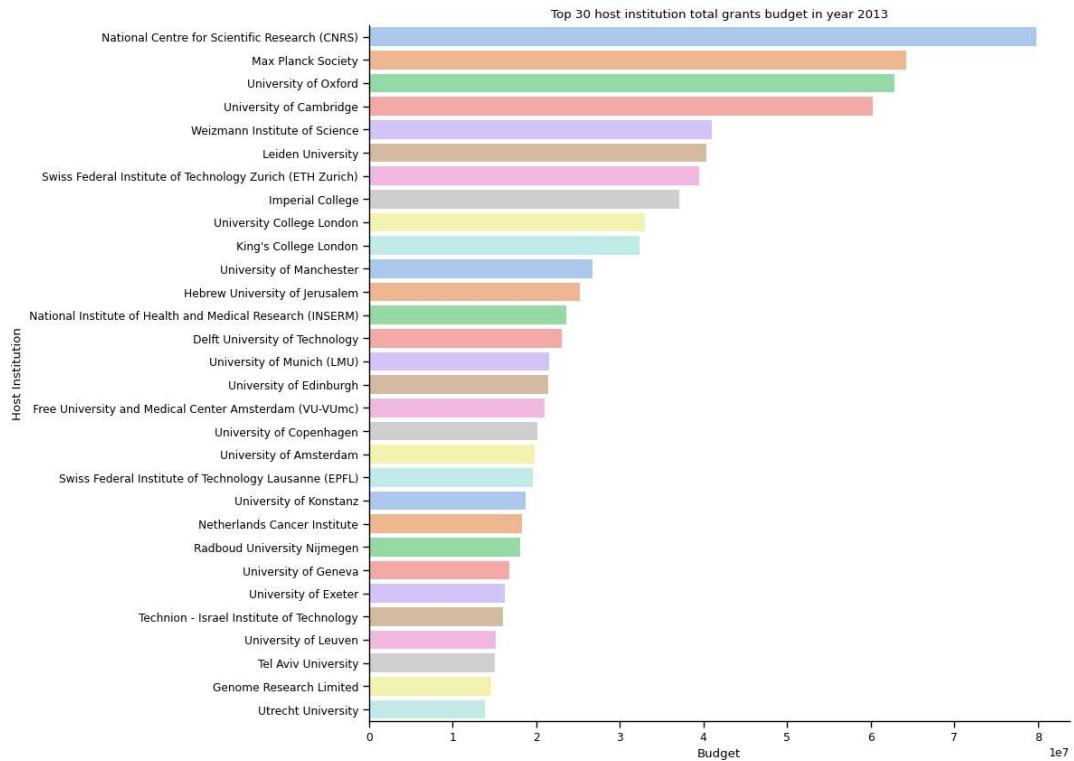
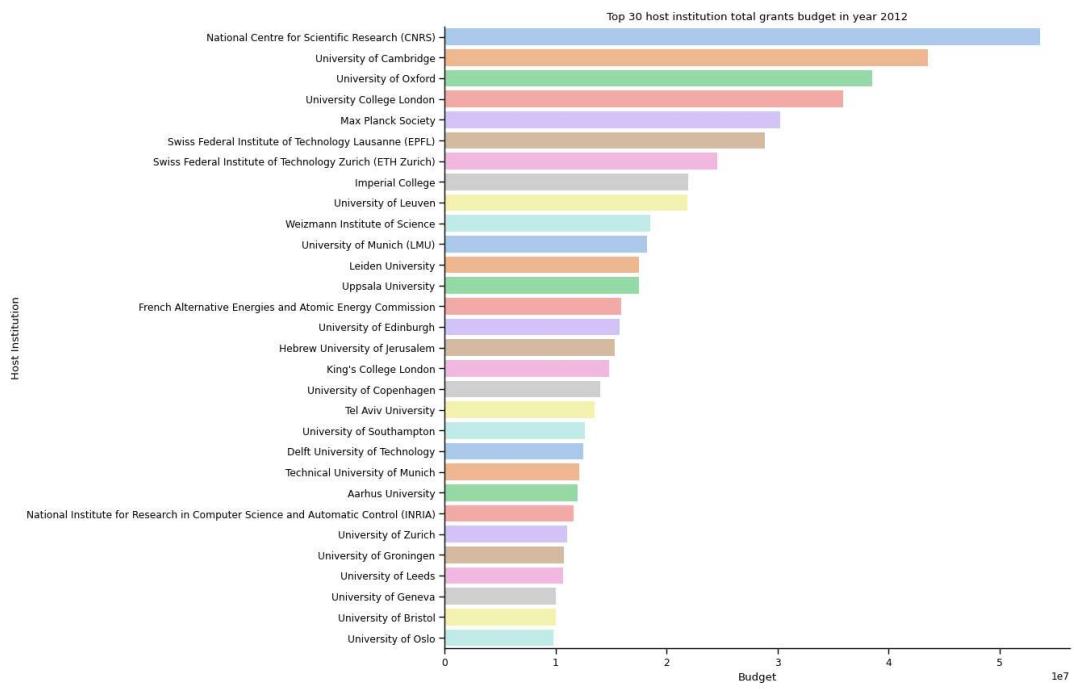
Supplementary figure 6: Distribution of grants per grant type along the years in the ERC dataset.

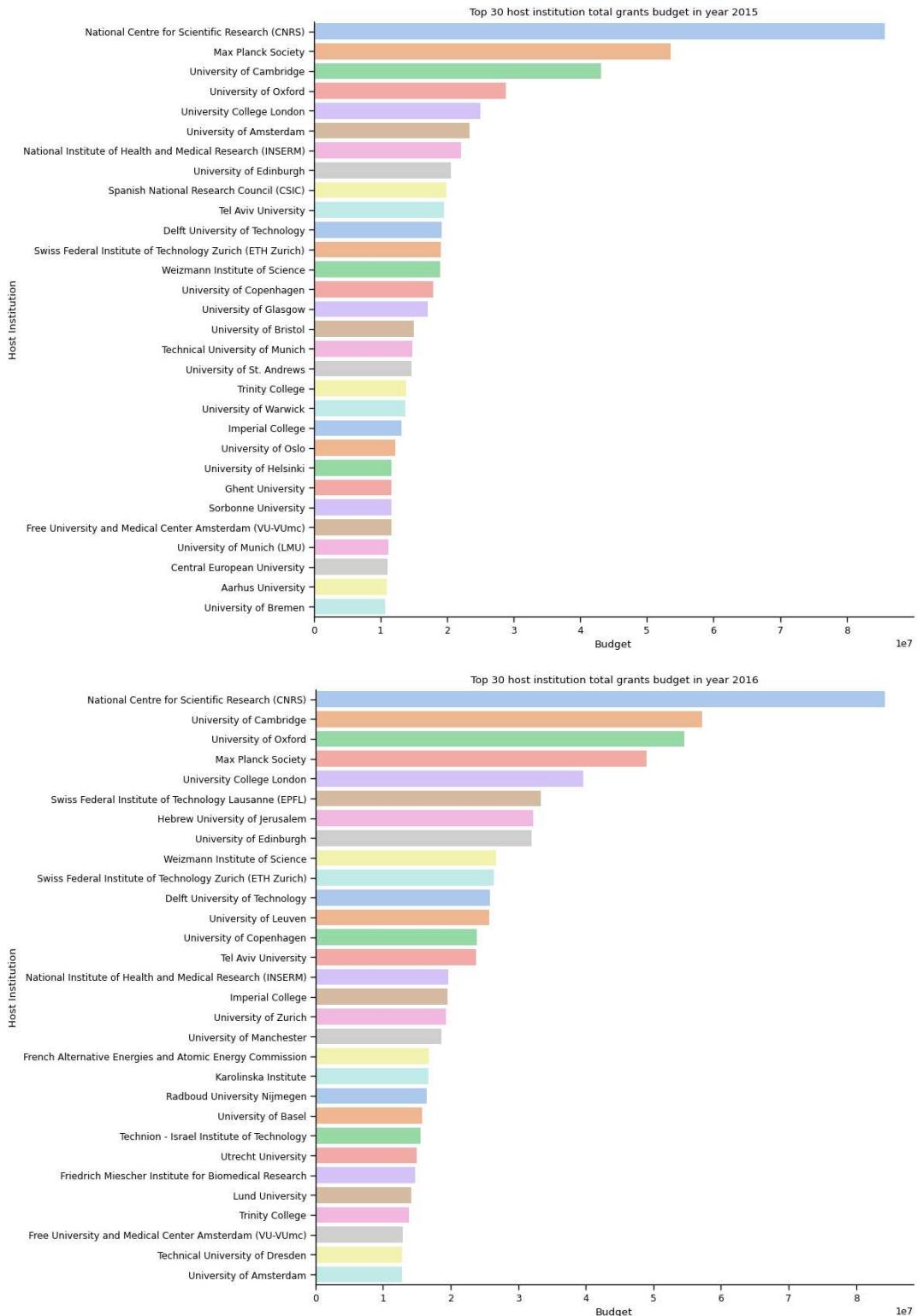


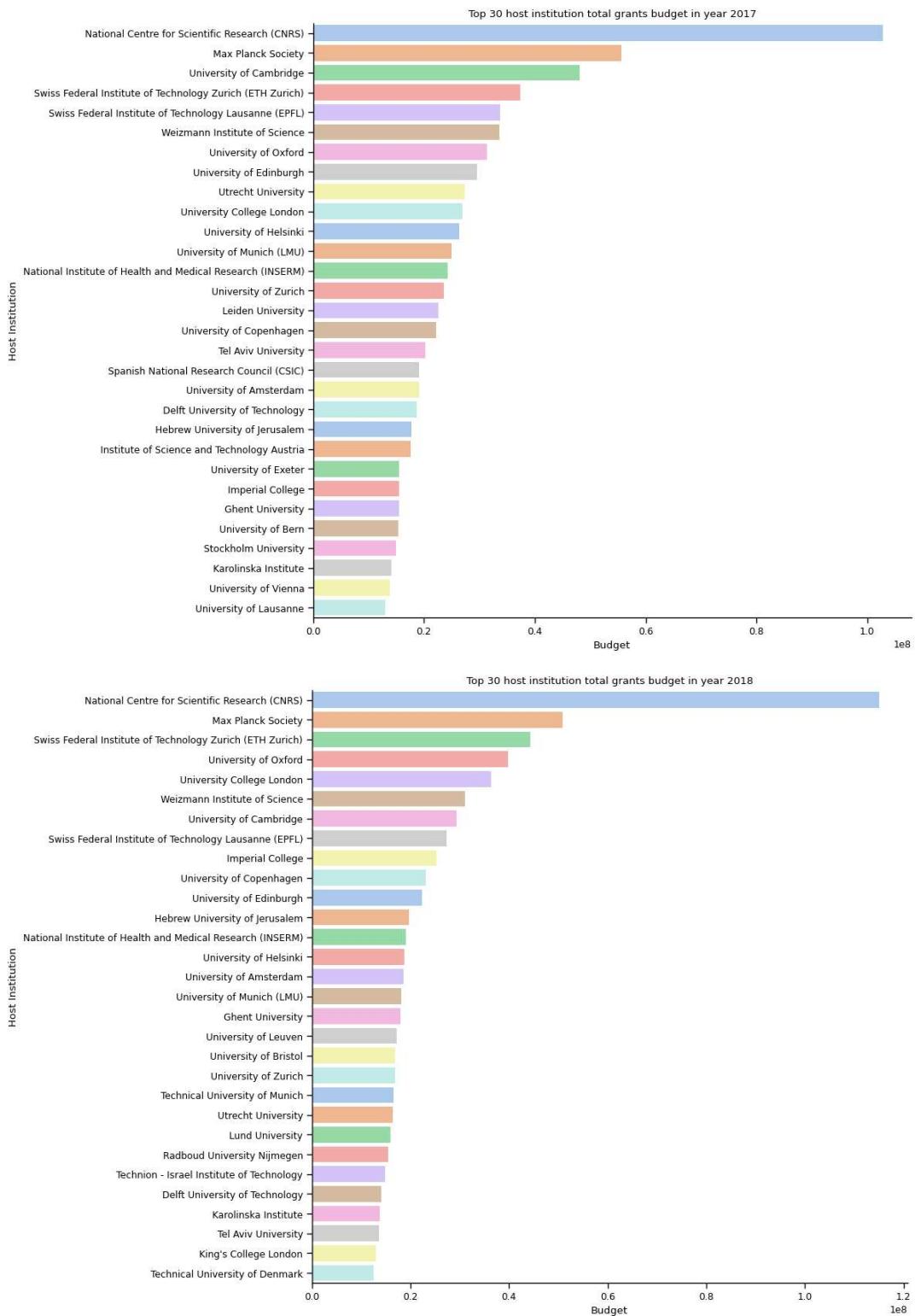
Supplementary figure 7: Distribution of grants budget per grant type along the years in the ERC dataset.

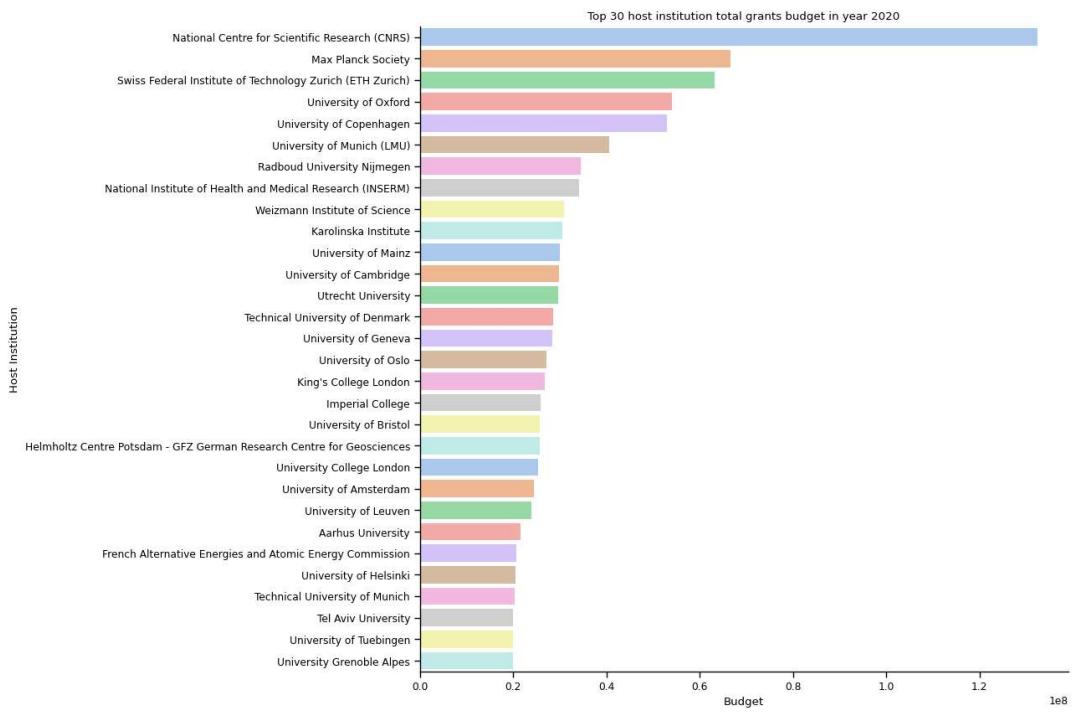
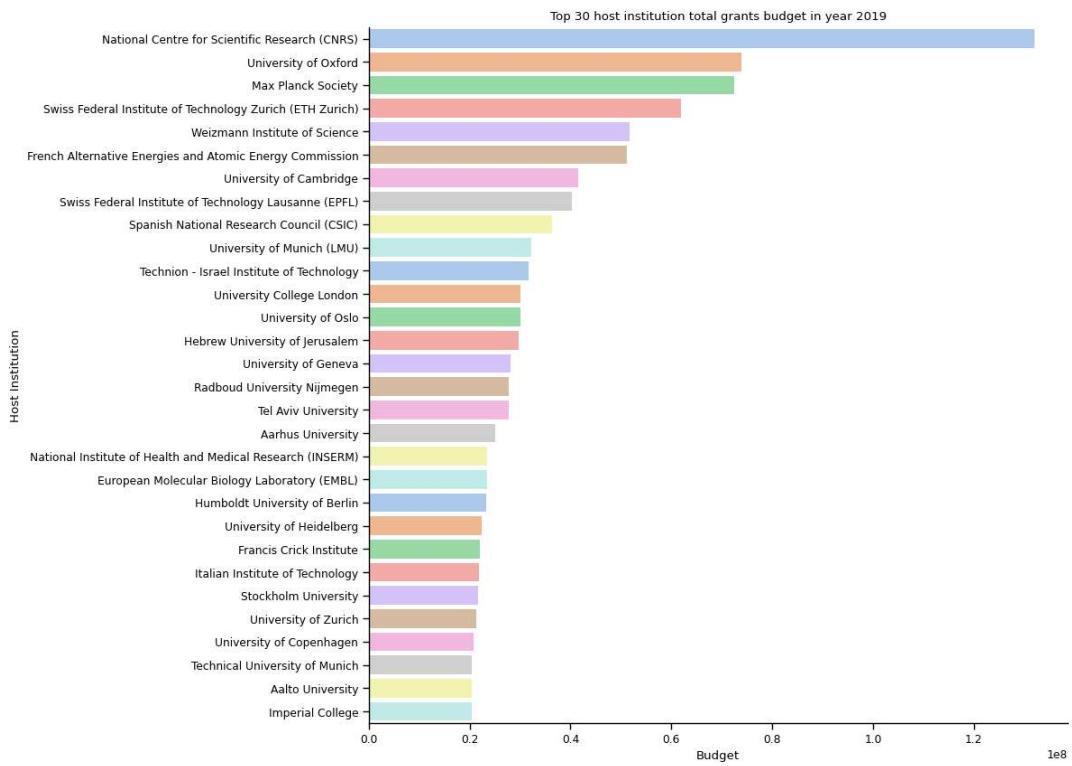


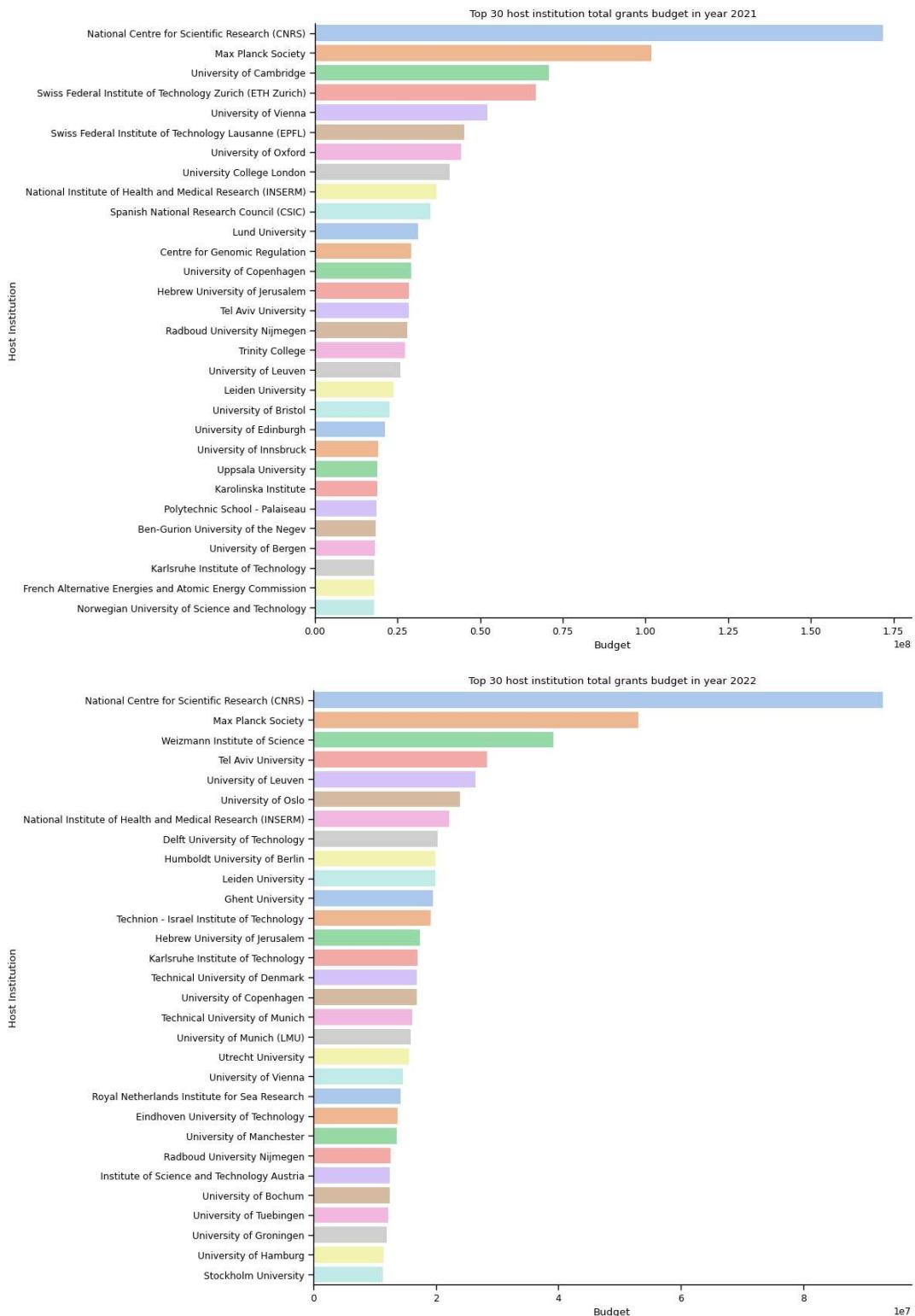


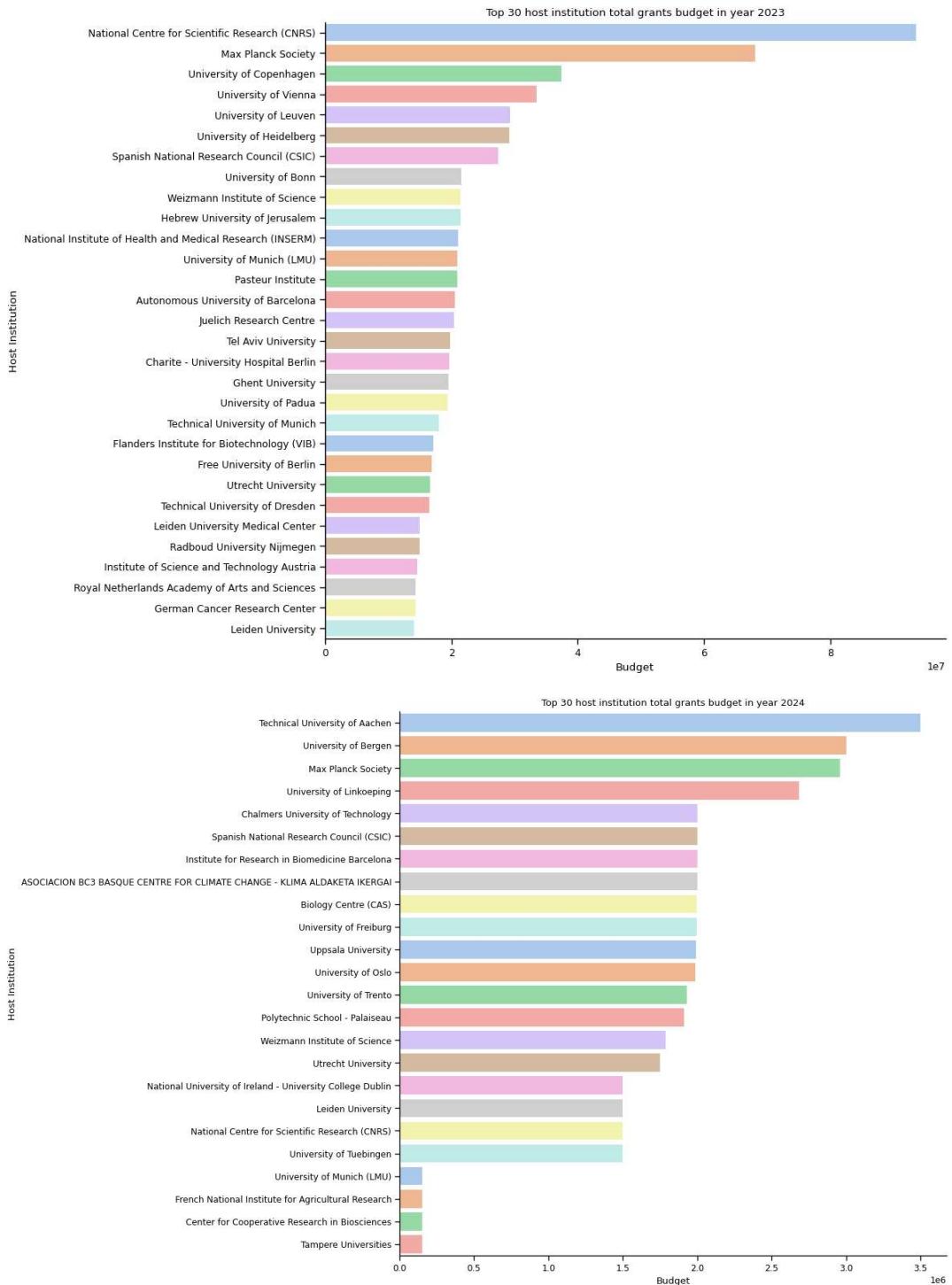






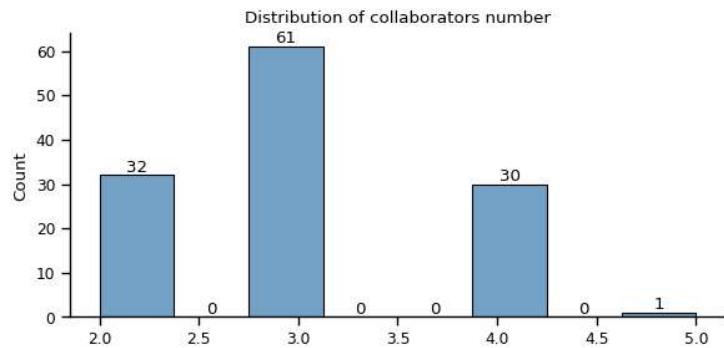




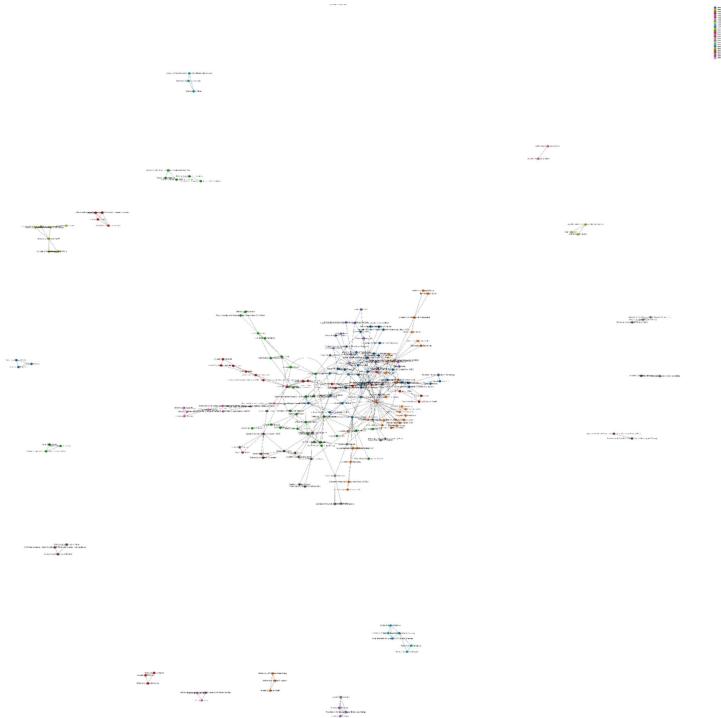


Supplementary figure 8: Distribution of grants budget per top 30 host institution along the years in the ERC dataset.

## Figures from network analysis

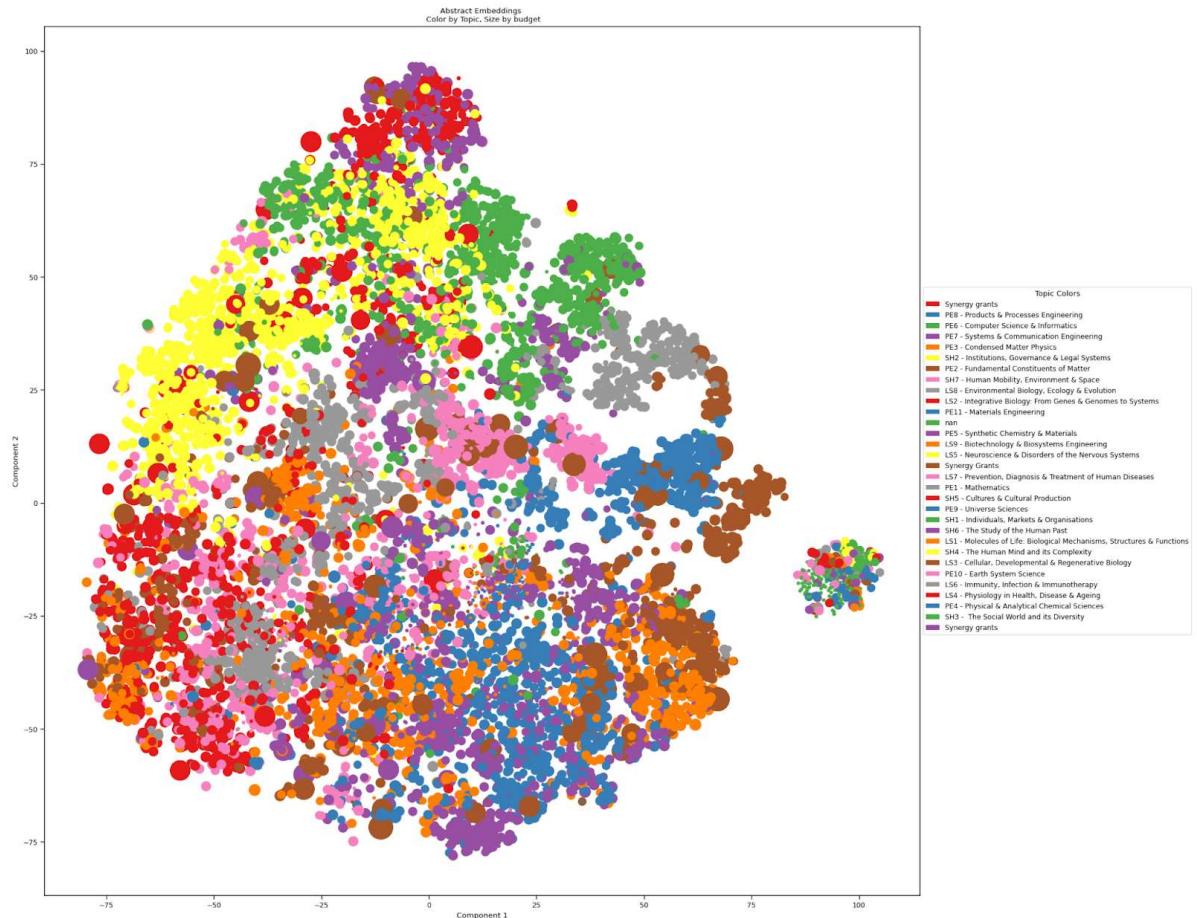


Supplementary figure 9: Distribution of PI's collaboration in the ERC dataset.

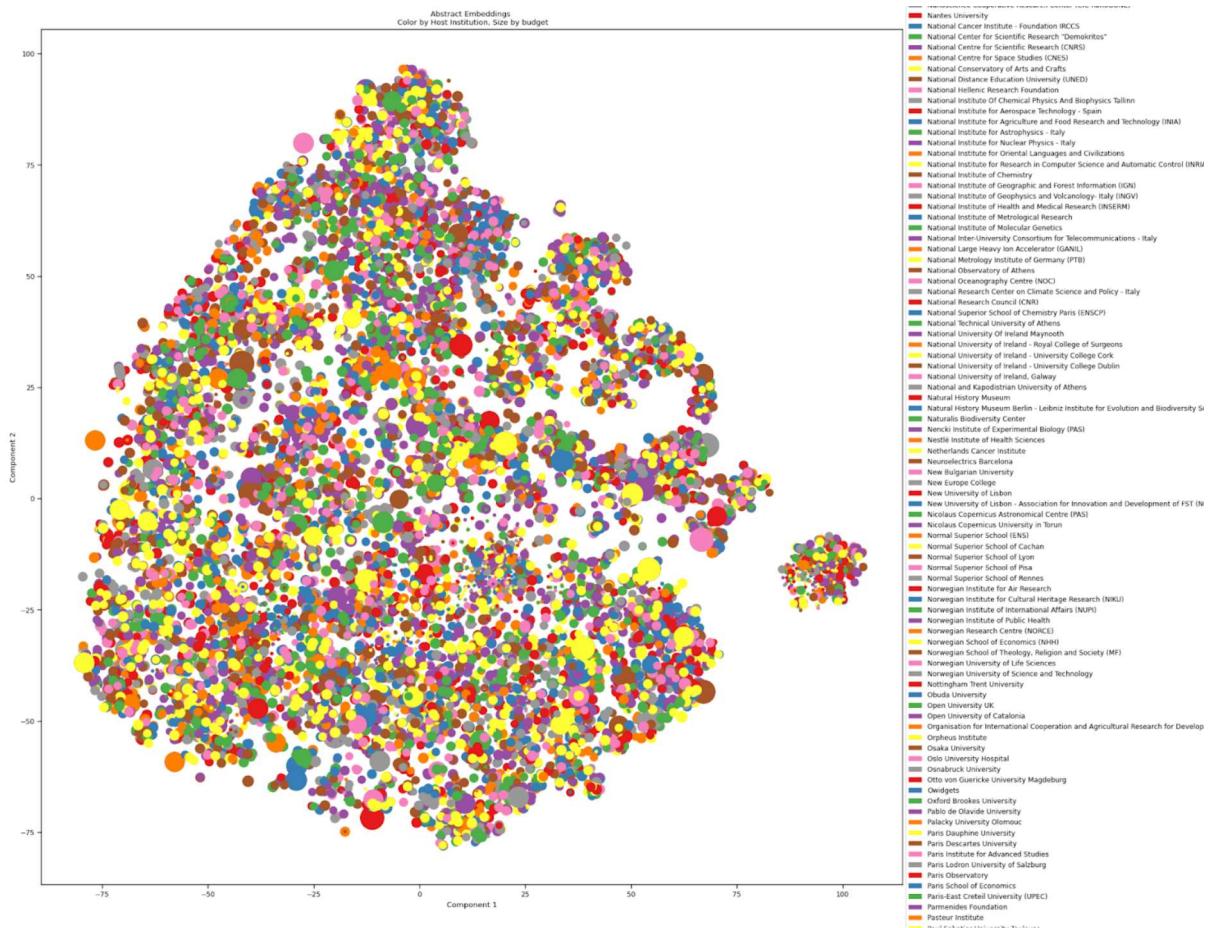


Supplementary figure 10: Graph communities in the ERC dataset.

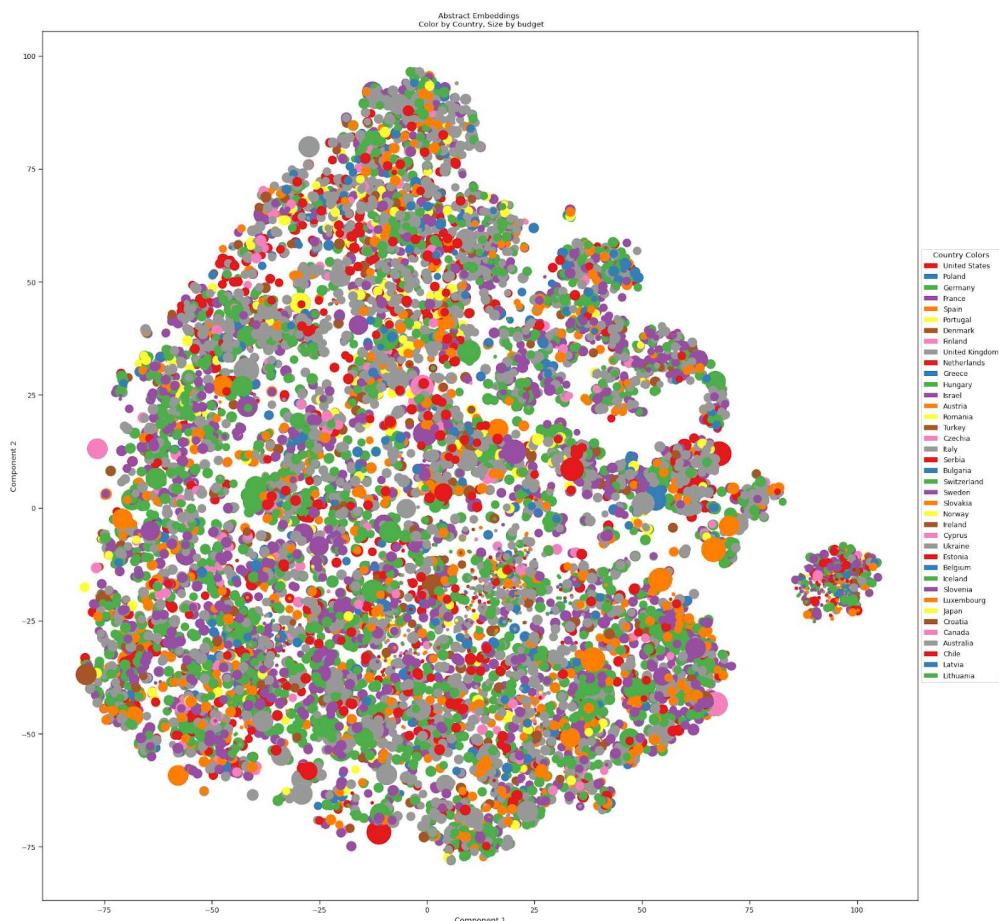
## Figures from NLP analysis



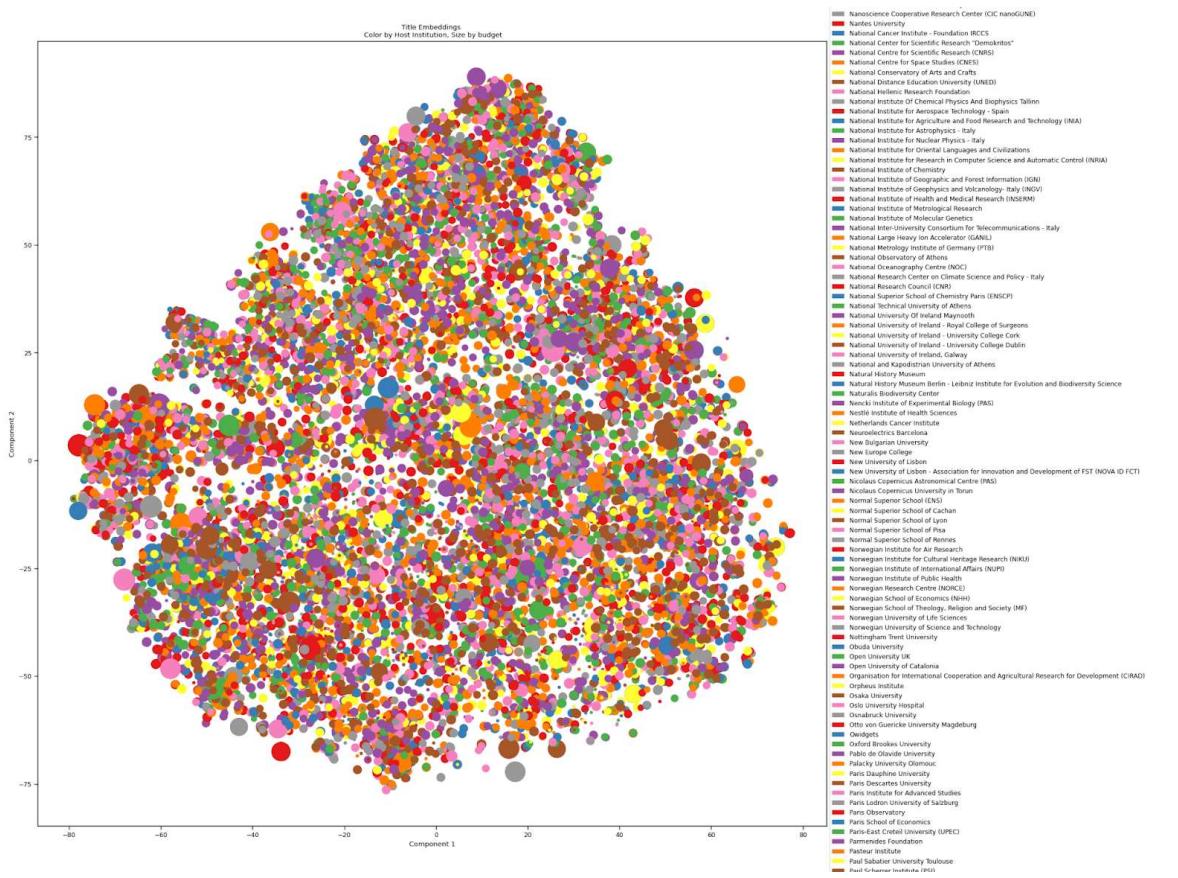
Supplementary figure 11: Grants abstract embeddings in relation to money (size) and topic (colour).



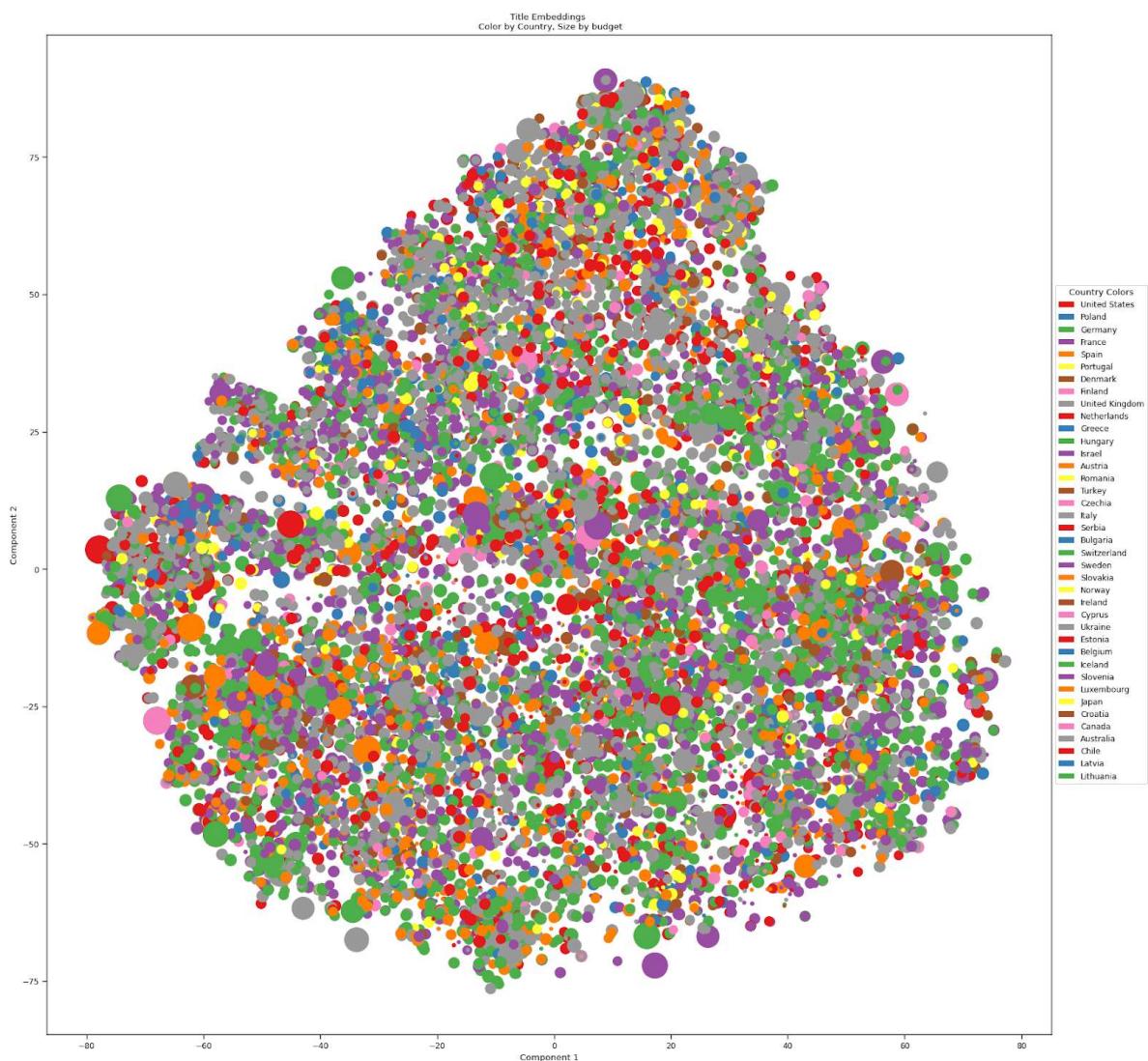
Supplementary figure 12: Grants abstract embeddings in relation to money (size) and university (colour). This is a partial image, for full image refer to the notebook attached.



Supplementary figure 13: Grants abstract embeddings in relation to money (size) and country (colour).



Supplementary figure 14: Grants title embeddings in relation to money (size) and university (colour).



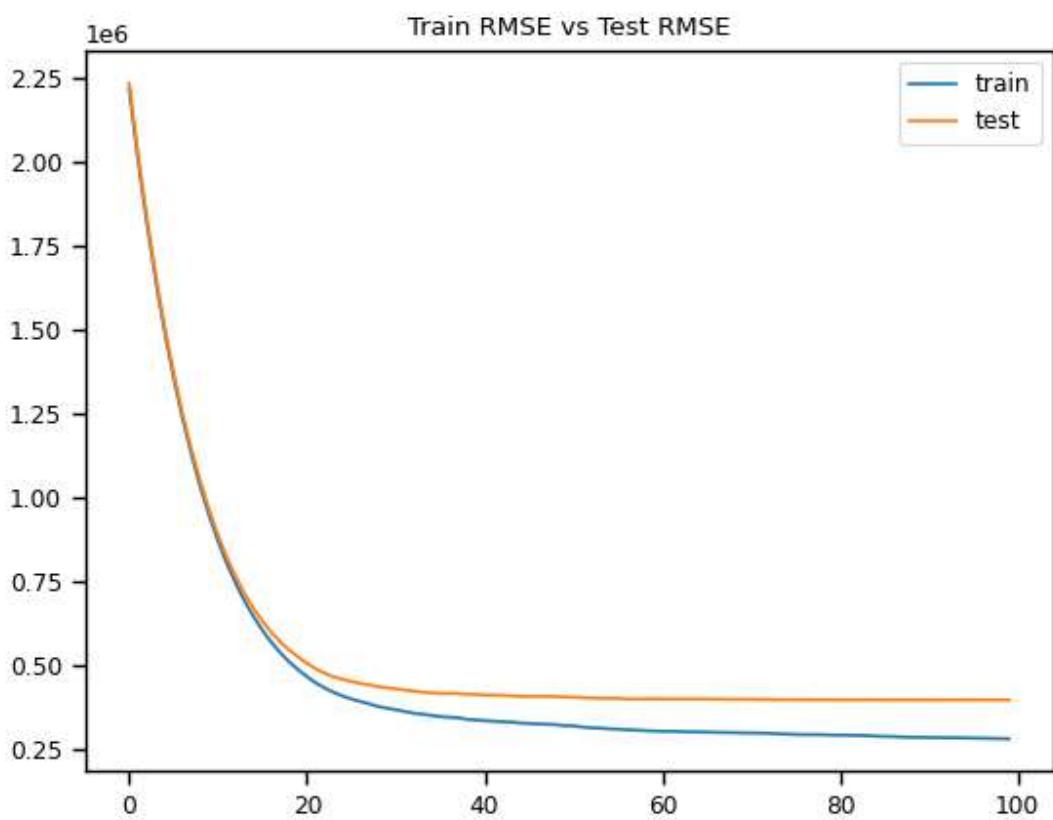
Supplementary figure 15: Grants title embeddings in relation to money (size) and country (colour).



Supplementary figure 16: WordCloud of grant titles in all the erc dataset.



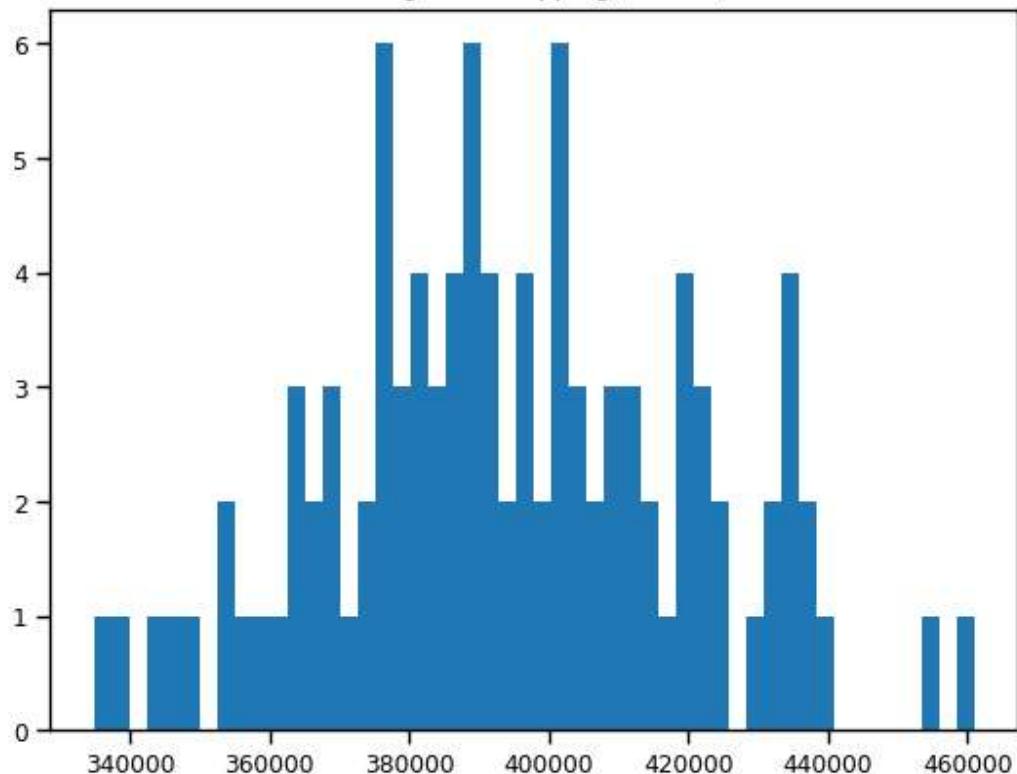
Supplementary figure 17: WordCloud of grant abstracts in all the erc dataset.



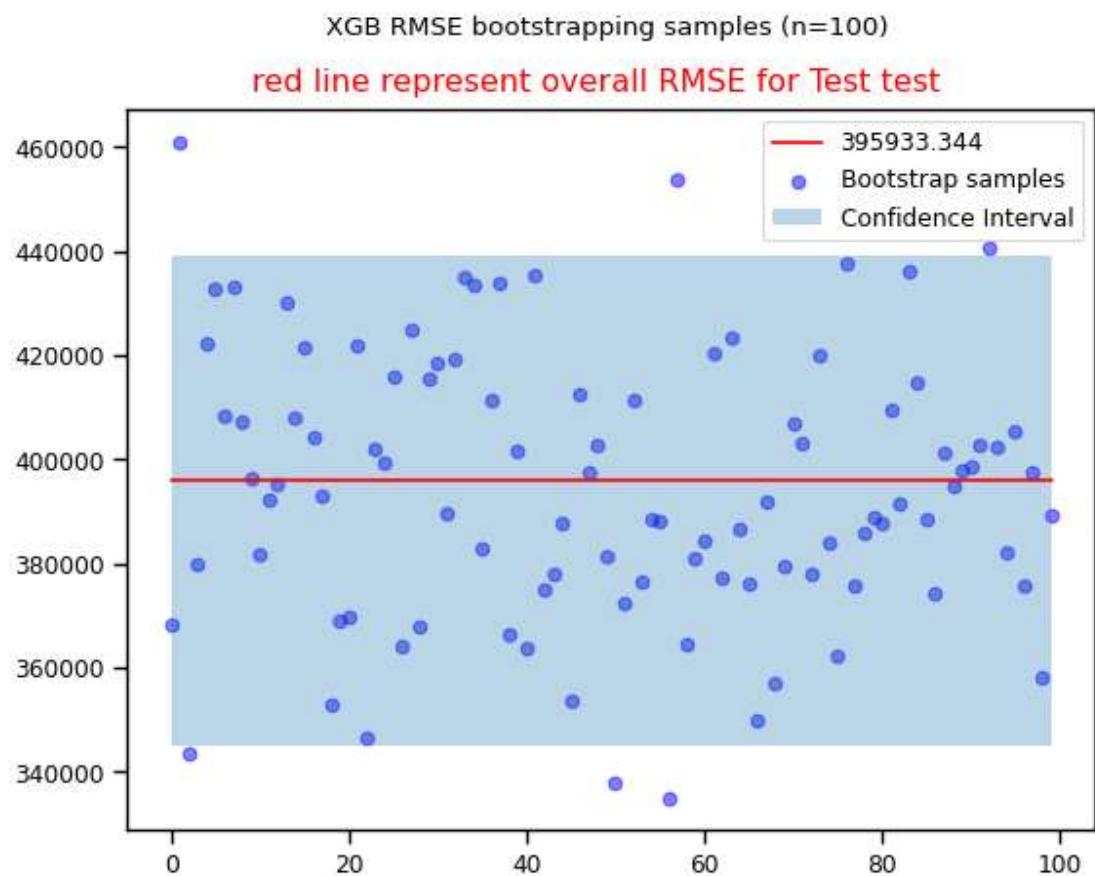
Supplementary figure 18: Train test mse during training.

XGB RMSE Uncertainty Measures  
95.0% CI is: [344858.340, 439079.731]  
standard\_error: 25899.635

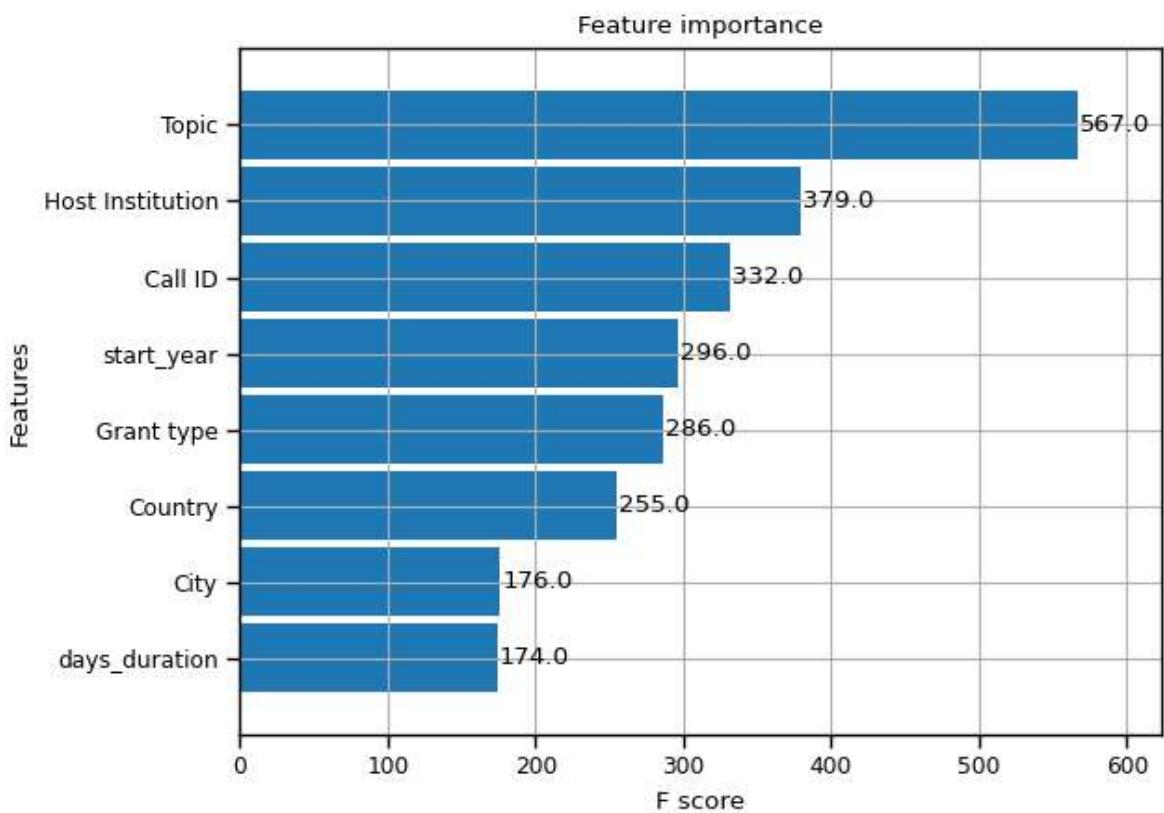
using bootstrapping (n=100)



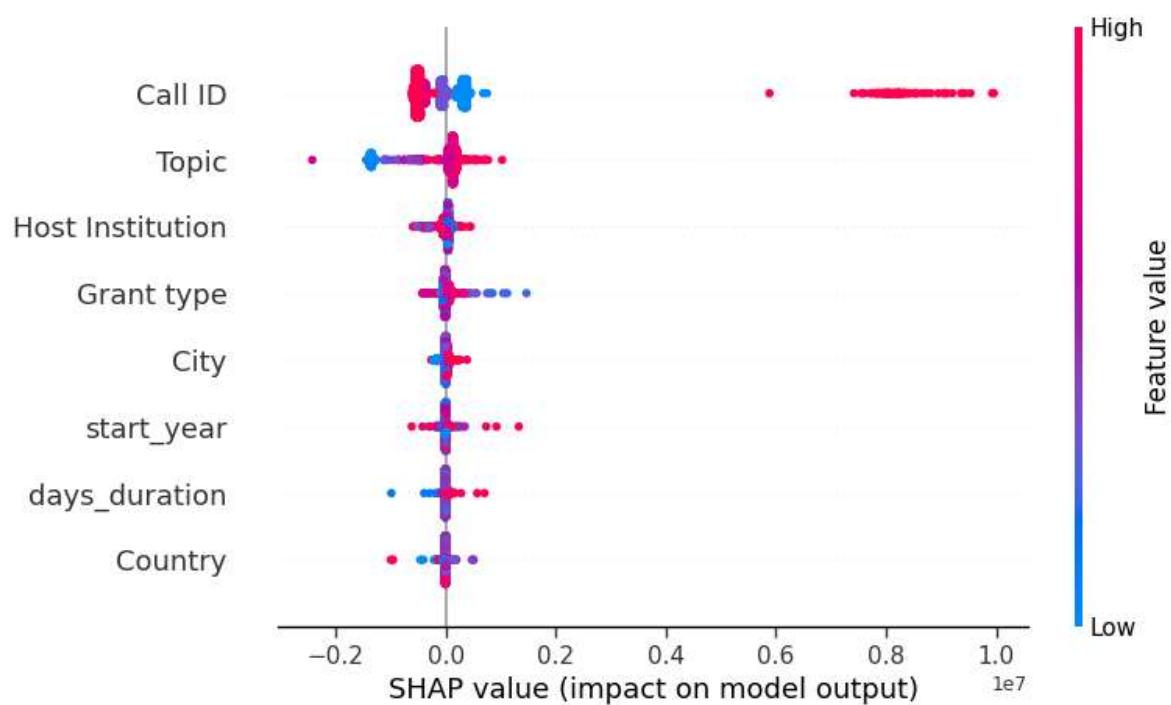
Supplementary figure 19: Uncertainty measure of XGBRegressor on the test set.



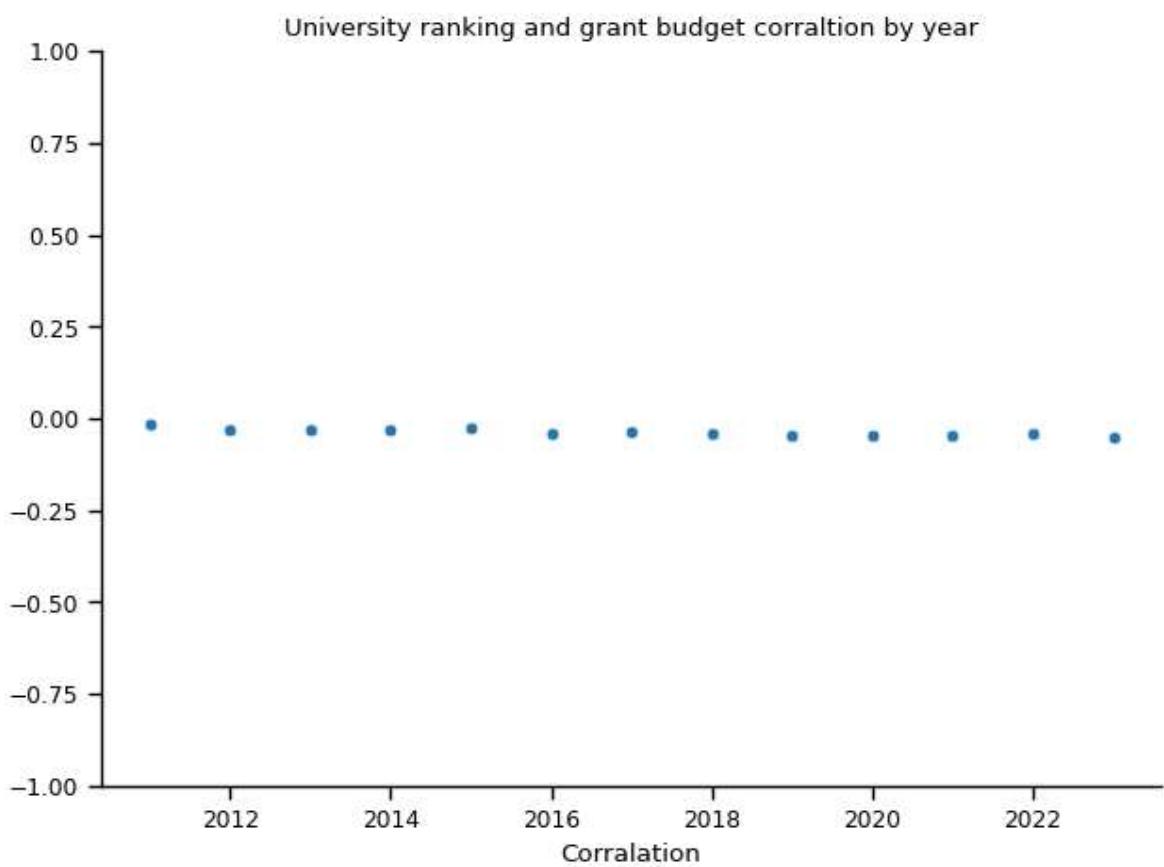
Supplementary figure 20: RMSE of XGBRegressor on the test set.



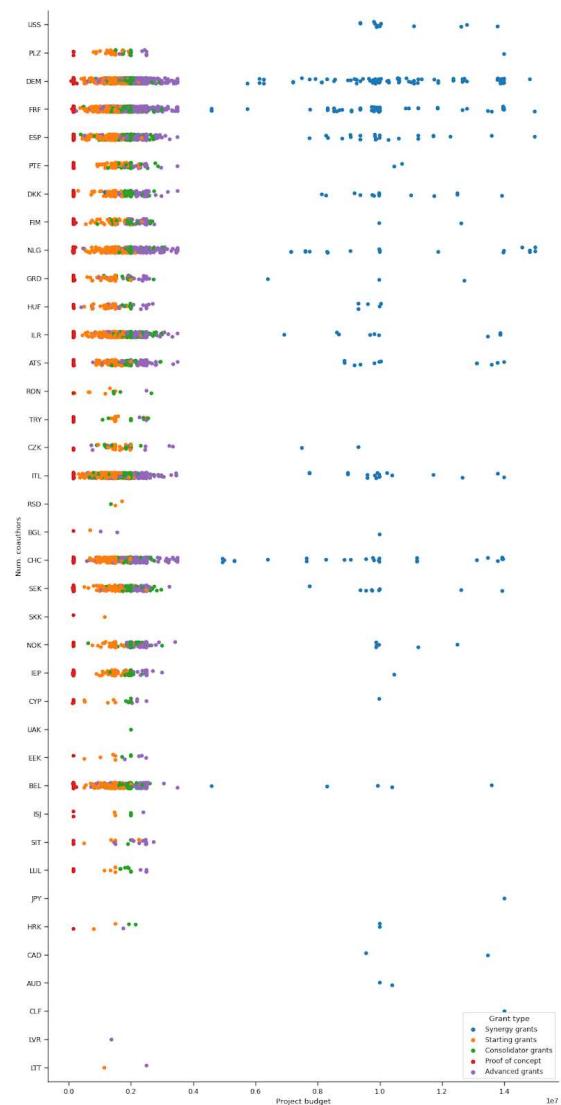
Supplementary figure 21: Feature importance of XGBRegressor on the test set.



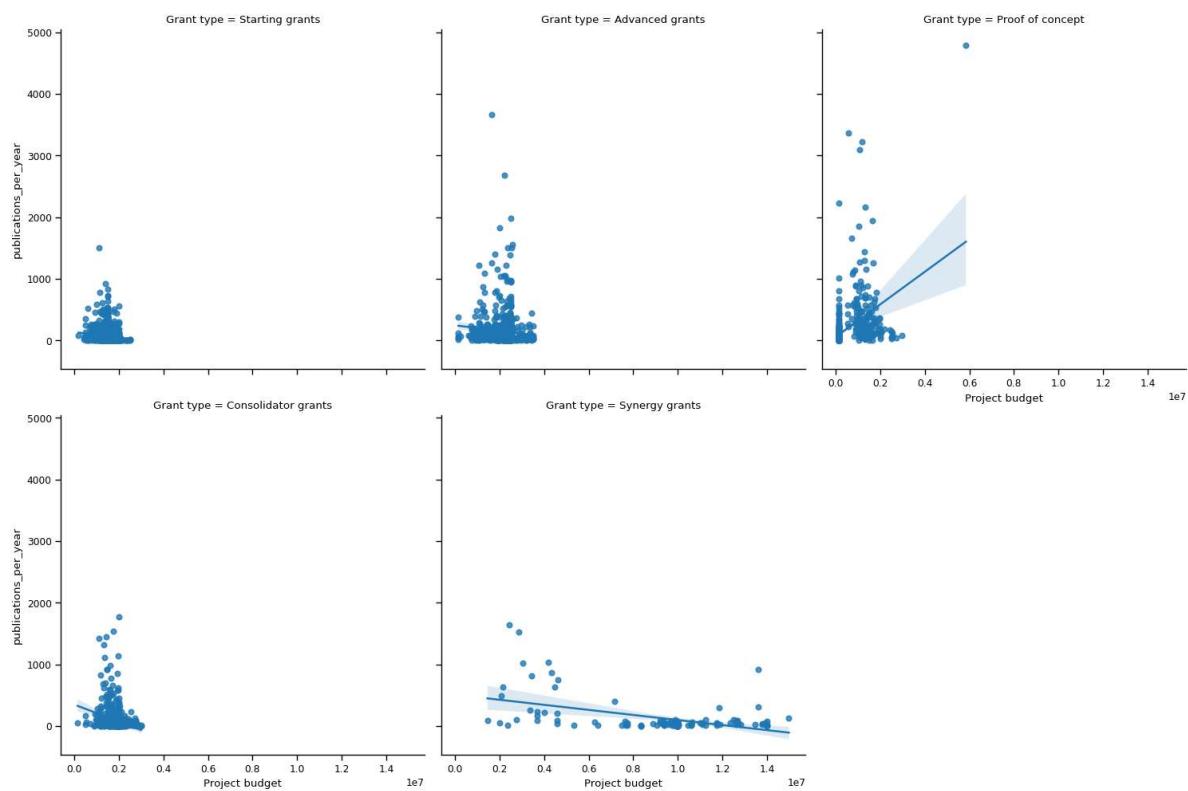
Supplementary figure 22: SHAP values of XGBRegressor on the test set.



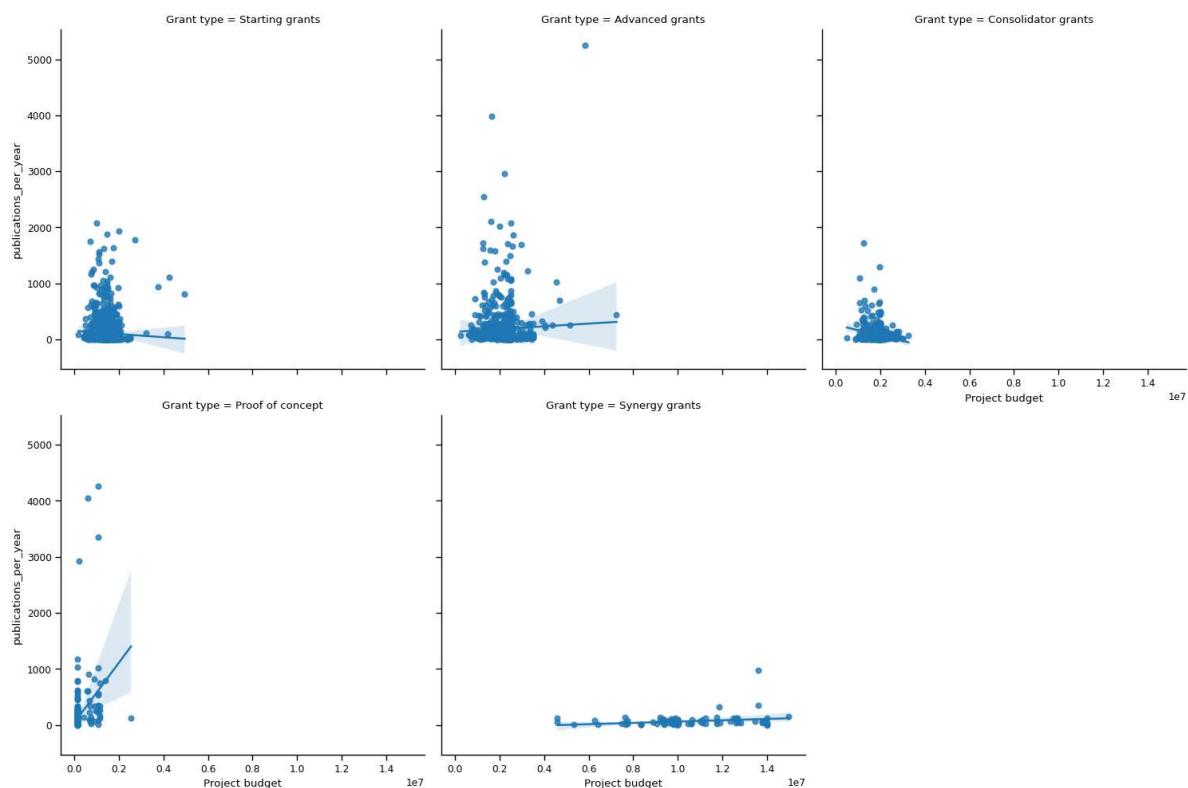
Supplementary figure 23: Correlation between university rank and budget per year.



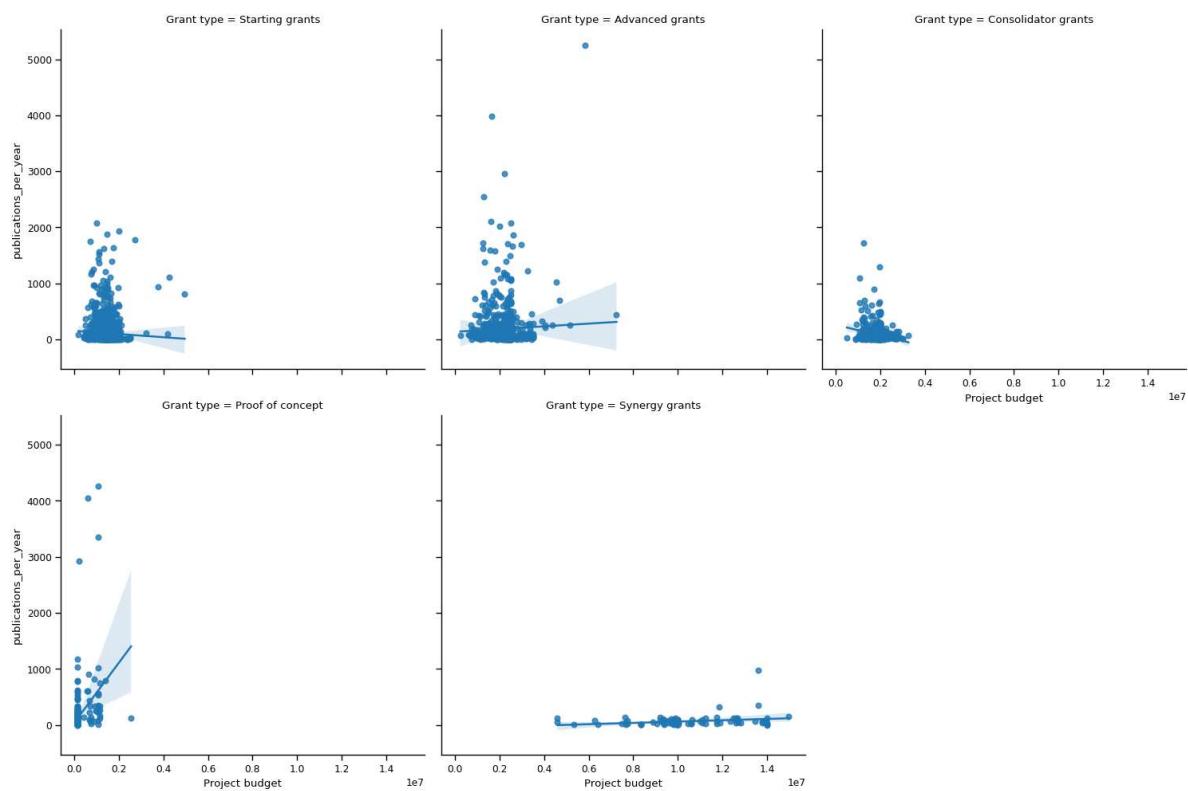
Supplementary figure 24: Correlation between currency and budget.



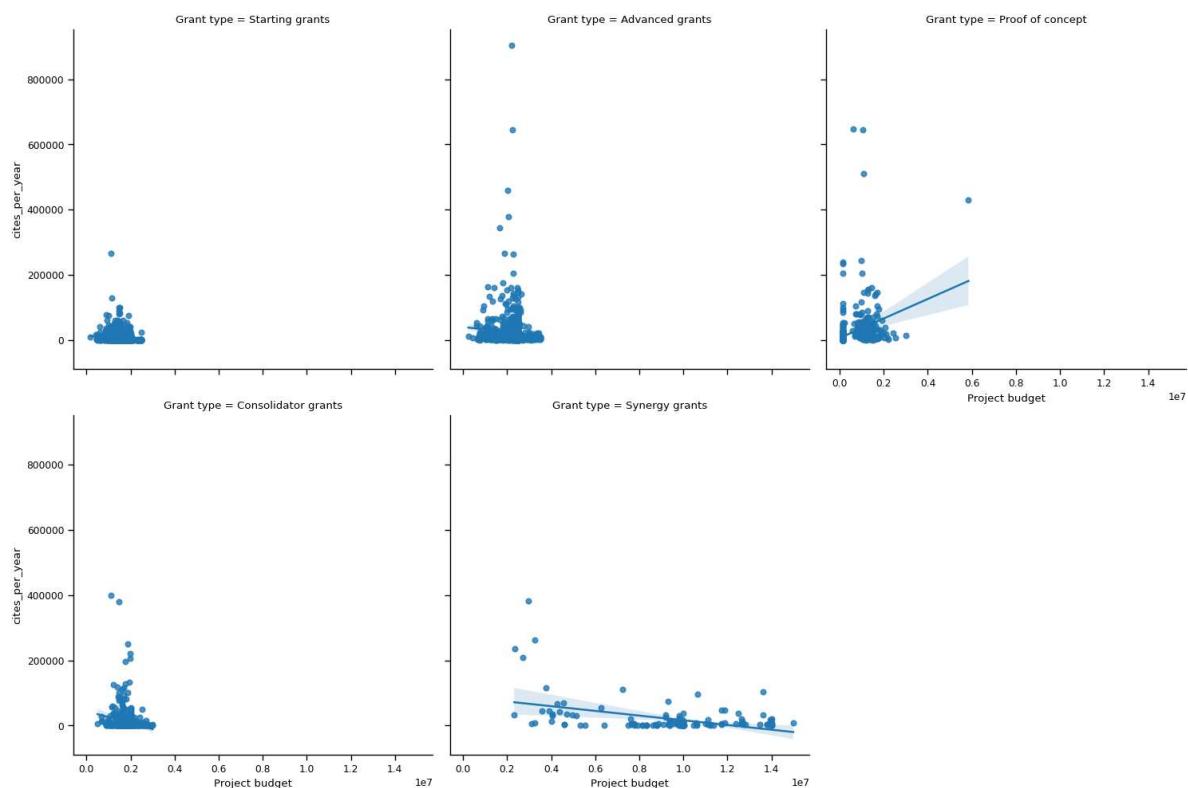
Supplementary figure 25: Correlation between grant budget (before getting grant) and PI number of publications



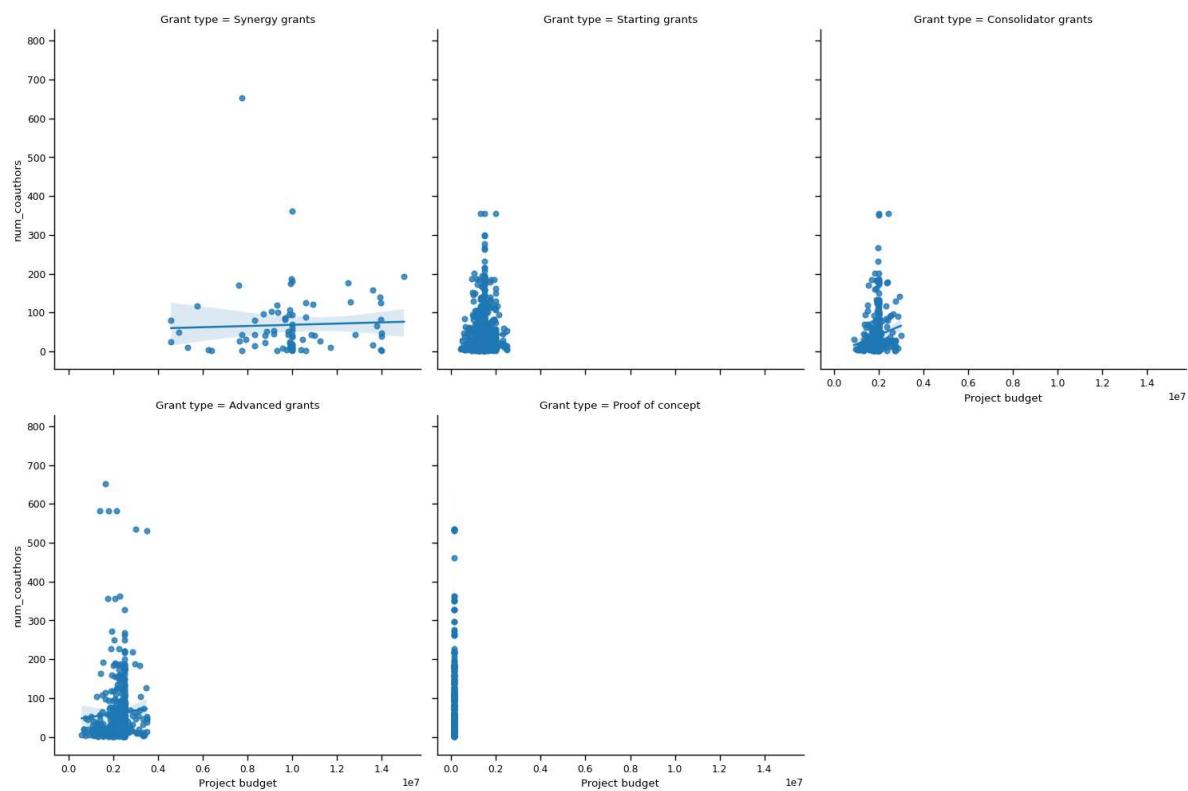
Supplementary figure 26: Correlation between grant budget (before after grant) and PI number of publications



**Supplementary figure 27: Correlation between grant budget (before getting grant) and PI number of citations**



Supplementary figure 28: Correlation between grant budget (after getting grant) and PI number of citations



Supplementary figure 29: Correlation between grant budget and PI number of coauthors