

# Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-03-16 17:32:26

PAGE 1

REFERENCE NO: 185

This contribution was submitted to the National Science Foundation as part of the NSF CI 2030 planning activity through an NSF Request for Information, [https://www.nsf.gov/publications/pub\\_summ.jsp?ods\\_key=nsf17031](https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf17031). Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>).

## Author Names & Affiliations

- Christopher Welch - Merck & Co., Inc.

## Contact Email Address (for NSF use only)

(Hidden)

## Research Domain, discipline, and sub-discipline

organic chemistry, analytical chemistry, separation science, high throughput experimentation, pharmaceuticals

## Title of Submission

Emerging need for Computational Resources for Precompetitive Collaborations on New Enabling Technologies for Pharma

## Abstract (maximum ~200 words).

Just a quick note to make the case for an emerging need for computational resources for current and projected precompetitive collaborations among pharma companies and academia in the area of new enabling technologies for synthetic chemistry, analytical chemistry, polymer chemistry and materials science as it pertains to pharmaceutical discovery, development and manufacturing. (I had received word from a colleague in academia that the response to this RFI from the chemistry community had been weak, so I just wanted to highlight a few items and issues of emerging importance to our field)

**Question 1** Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Big data science is coming slowly to the field of chemistry, but in our labs recent research has convincingly shown the power of this approach.

However, our experience is also showing that many if not most data sets in the chemistry field tend to be fragmentary and strongly biased toward successful results - limiting their value in developing models that correlate molecular structure with properties like solubility, reactivity, chromatographic retention, MS ionizability, viscosity, etc., etc. In addition, current data sets are often 'siloed' within companies and institutions, limiting availability for sharing.

this recent paper from our group discusses this problem in some detail, for a specific area of analytical chemistry....  
<http://www.sciencedirect.com/science/article/pii/S0021967316306732>

# Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-03-16 17:32:26

PAGE 2

REFERENCE NO: 185

---

(and we will soon submit a publication dealing with chemical reactivity databases)

We recently created a cross-pharma consortium dedicated to precompetitive collaborations on new enabling technologies ([www.etconsortium.org](http://www.etconsortium.org)). Many discussions within this group envision future sharing across organizations to amass data sets with sufficient breadth and scope to enable successful data mining approaches.... as much of this information is proprietary, amalgamating data in ways that can enable the extraction of valuable structure-activity relationships without divulging exact structures will be needed. Ideally, a shared resource in the middle would be available for doing this in a secure and trustworthy manner - and perhaps this could be a good fit for an NSF funded computational facility.

In addition, the field of chemistry could benefit from the input of big data analysis experts from other fields as we attack this problem

finally, we are becoming pretty strongly convinced that, given the sparseness and extreme success bias in much of our data, we may require de novo collection of much of the needed experimental information in collaborative research projects utilizing automation and high throughput experimentation. The availability of a shared space in the middle for securely warehousing data, providing consultation on how best to set this up and help with expert guidance for data mining would be extremely valuable, and we would like to suggest that these would be valuable capabilities to build into envisioned future NSF cyber capabilities.

**Question 2** Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

thanks - see q1

**Question 3** Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

thanks - see q1

## Consent Statement

- "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."
-