A competition for computational toxicity forecasting

Evan Miyazono & Anish Goel

Summary

Start a benchmarking competition for rapidly and accurately predicting the kind and degree of chemical hazards.

Significance

AI advancements in biochemistry present a significant challenge: while they enable the rapid development of new chemicals and processes, they also outpace the current regulatory and review frameworks, including toxicity assessment methods. This growing gap between action and oversight will increase the risks of poorly understood chemical hazards entering foods, products, and the environment.

In response, we propose the Critical Assessment of Toxicity Prediction (CATP) competition. CATP is designed to provide three things: (1) a comprehensive toxicity forecasting dataset to be leveraged by AI researchers to advance in-silico risk forecasting, (2) a clear benchmark for evaluation of competitors, and (3) an ongoing competition deadline to stimulate regular progress. By fostering a collaborative and competitive environment, similar to the successful Critical Assessment of Structure Prediction (CASP) experiment, CATP will be an open-source, benchmarking competition.

Rather than simply recreating the success of CASP and AlphaFold, we argue that toxicity forecasting is the most important extension of AlphaFold's success in leveraging AI to advance our understanding of biology. (Specifically, Alphafold combined "bottom-up" training data on protein structure with "top-down" understanding combining mutation correlations across species with spatial relationships. Analogously, we expect that known toxicity information will provide bottom-up data, while the AlphaFold Protein Structure Database can provide top-down intuition of bioactivity.) Through this initiative, we aim to stimulate the creation of qualitatively faster and more accurate tools for toxicity prediction, ensuring that the rapid advancements that AI holds for the future of chemical synthesis and pharmacology development are matched by equally rapid and reliable methods of safe chemical review.

Proposal

Very Rough Timeline: (I'd love to make this a gantt chart, but opted to keep this doc under 2 pages)

- **Now - greenlight:** Iterate on proposal, identify potential hires and collaborators, secure funding
- **Months 0-2:** Start assembling a team, begin assembling dataset(s), generate scoring metric proposals
- **Months 2-4:** Finish hiring core team, continue collecting & validating data, peer review scoring metrics
- **Months 4-6:** Finish data collection, announce competition, recruit early competitors
- **Months 6-8:** Start organizing competition, go/no-go on foundation setup for recurring competition
- **Months 8-10**: Run competition

Key Milestones: **(1)** Successfully integrate multiple existing public (and ideally some private) datasets. **(2)** Collaborators from academia, industry, and regulators are strong advocates for a potential scoring metric (and perhaps a "success" threshold for the competition). **(3)** Competition launch w/ participation from at least 1 of the top AI labs. **(4)** A competitor beats the threshold set in milestone (2).

Key Technical Risks: **(1)** Integrating and standardizing vast amounts of data from diverse sources like Tox21, eTOX, and PubChem into a useful training dataset. **(2)** Acquiring enough data to be confident models will generalize. **(3)** Developing input/output pairs that both are useful and match available training/test data. **(4)** Maintaining the robustness and scalability of the digital infrastructure to support a critical mass of participants.

Founding Team and Talent

We believe that to execute this project effectively, it is imperative to have experts spanning from AI, computational toxicology, digital infrastructure, and those with experience organizing open-science competitions (e.g., X-Prize, CASP, Tox21). As a result, we propose the following structure and background:

Core Founding Members:

- *CEO/Product Lead*: strong experience in computational bio-chem and passionate about safe advancement of AI. Ideally, has experience running a leading CASP participating team.
- *Lead biochemist*: relevant background in computational toxicology
- *Lead data manager/curator:* experience with biological data (perhaps from Tox21), ideally data for training machine learning systems
- *Competition architect:* likely someone from XPRIZE, or other similar challenges that can assist with event organizing, digital infrastructure, and library science

This group could grow to include greater data stewardship or marketing, depending on availability of funding.

Advisory Board:

- Experienced researcher(s) from DeepMind who contributed to AlphaFold's development on the ML & biochemistry side
- Former organizer(s) from the CASP competition
- Stakeholder(s) from a relevant gov agency (e.g. NIH, NSF, DOD, EPA, EMBL) or program (e.g. Tox21)

There are likely other areas from which advisors could bring relevant experience and world-models.

Budget

As a reference, the CASP15 (2022) team consisted of 5 core members (organizing committee). As a very rough estimate, assigning $250,000 USD salary+overhead to each core member, this is $1.25M/yr for staff. I'd estimate $250k for additional overhead (design, infra, communications, and organizing a remote event), bringing the total to roughly $1.5M for the first year, and potentially less for future years (especially if the event could eventually be funded by the NIH, as CASP is).

Why this can't be done in academia/industry alone

University labs do not seem to be incentivized to build a dataset of this scale as it requires significant efforts toward logistics and organization. Providing ongoing hosting of the dataset or hosting a competition would be out of the question, so impact would be limited to what can be achieved by a single lab.

If an industry actor set out to create this competition, creating an open dataset and competition would likely be yielding a competitive advantage in the development of proprietary technologies. It's also possible that improving forecasting toxicity could be seen by industry players as likely to create a net increase in oversight bureaucracy of drug development or manufacturing and materials processing, as common (by)products are found to be harmful.

Presumably government science funders could do this alone, but may not be fast enough to respond or capable of taking the risk of setting up such an event, though the NIH or NSF could presumably sustain the competition once it was derisked.

Sharing

This is a public document. If you'd like to support this work, feel free to schedule time with me.

Appendix 1: What data is needed, and motivating the feasibility of the scoring metric

To prime an intuition pump, imagine a vector space with one dimension corresponding to each molecule in the cell. The effect that any therapeutic has on the biological system is then expressible as one or more vectors describing something like interaction strength or binding energy. In practice a combination of many such chemical descriptors might be used. (The natural metabolic cycles would then be the blocks in the roughly block-diagonal self-interaction matrix.)

Two datasets would then be needed:

- Data listing chemical descriptors of various small molecules.
- Data mapping chemical descriptors to relevant toxicology factors, like LD50.

The need for an information-dense metric such as this arises from the fact that a simple binary or scalar estimate of toxicity is not as verifiable. Ideally, predicted interactions between a test molecule and any component from life should be able to be reduced to a quantity that can be physically measurable or perhaps simulated with non-ML-based techniques to allow predictions to be efficiently spot-checked as the database of biological interactions grows.

Both of these datasets are difficult and costly to generate, as they're currently limited to in vitro and model measurements. Additionally, as chemical descriptors might be guarded proprietary information and there are incentives to release toxicology information as quickly as possible. To run a competition effectively, it could be possible to have a rolling deadline rather than a fixed event (as CASP does), where newly released measurements are embargoed only long enough to run all the models.

I wouldn't expect that this competition would yield results as comprehensive as AlphaFold 2's "88% of AlphaFold 2's predictions had an RMS deviation of less than 4Å for the set of overlapped C-alpha atoms", but would instead likely achieve accurate forecasts for <<80%, based simply on the size of the parameter space and the amount of data.

Appendix 2: Mitigating risks of this technology

- When predicting interactions with cells, there is more risk in simulating larger, more complex molecules that have the potential to self-replicate. ⇒ Therefore, training and test data should be limited to small molecules
- Any accelerated modeling of physical systems is likely to be dual use. ⇒ Participants should be required to pledge (or legally commit) to not making model weights open source; training data could be shared contingent on such a commitment, and training data could be water-marked to deter leaks.

Appendix 3: Motivating the toxicity metric as a biointeraction specification language

Part of my motivation for proposing this competition is that I'm advancing an architecture to govern AI based on specification languages at Atlas Computing. I believe that the bioactivity vectors and AI tools to predict bioactivity in silico from a chemical formula will be necessary for humans to constrain and oversee the behavior of advanced AI in the same way that these toxicology tools are needed for regulators to oversee corporations.