Open
Molecular
Software
Foundation

POSE: Phase II: Building open source ecosystems in molecular sciences through collaboration and technology

Project description of the NSF POSE grant proposal submitted in Oct 2022
NSF POSE solicitation: https://www.nsf.gov/pubs/2022/nsf22572/nsf22572.htm

PROJECT SUMMARY

Overview

Computational modeling has become an indispensable tool in modern research, including chemistry and biology. Development of computational frameworks supporting molecular engineering applications, such as a rational design of molecules with desired properties, is of particular interest across industries. Both academic and industry researchers are increasingly recognizing the value of high-quality open source software and pre-competitive collaboration efforts – new organizational and legal frameworks are needed to efficiently establish and manage such multi-institutional cooperations, as are new mechanisms for building and maintaining shared infrastructure. With the Open Molecular Software Foundation (OMSF), we seek to address this gap and create a collaborative, cost-effective organization to support an ecosystem of open source projects in molecular sciences, build expertise around managing and governing these projects, explore pathways to sustainability, and accelerate innovation by sharing our knowledge and tools under open licenses. Ultimately, our goal is to grow and support our **ecosystem as a distributed network of autonomous communities** connected via shared infrastructure, processes, and values.

Intellectual Merit

OMSF has 4 different projects under its umbrella and close ties with several adjacent projects, giving it a good initial grasp of the common challenges and potential solutions applicable to distributed, collaborative projects across the molecular software space. In the grant period, we will: 1) create “molecular sciences OSE playbook” – a combination of legal templates, onboarding guides and training materials for users and contributors, and governance and management processes for projects; 2) build technical solutions to improve quality and security of our projects, such as reusable kit frameworks, project cookiecutters, automated testing on different hardware platforms, and deployment via distribution systems; 3) strengthen OMSF’s position as a neutral ground where different stakeholders can come together to discuss and contribute to common infrastructure, securing long term sustainability and relevance of our OSE in the shifting landscapes of research needs and technology advances. All materials will be released under permissive licenses. We will build an active community around open source tools and best practices in molecular sciences and create cooperative mechanisms for continuous improvement and development of new technologies.

Broader Impact

The ability to perform efficient and productive research is contingent on the access to robust, reliable and well-documented research software regularly benchmarked for performance and accuracy. Biotech and chemical industries are becoming more reliant on molecular modeling in their R&D workflows, and the interest further grows with the recent advances of AI technologies. With OMSF, we are building a mechanism for rapid progression of the state-of-the-art computational methods in molecular modeling into stable, customizable and interoperable software to further expand research capabilities and improve outcomes. In addition, training and validating computational models require large datasets and are computationally very expensive, especially AI models, making this technology out of reach for less-resourced researchers and companies. We aim to broadly enable innovation by coordination of efforts, pooling resources and making resulting materials widely available for reuse and further development. Additionally, we will enable community oversight and control of future developments by adopting a community governance approach and transparency around these tools without sacrificing progress and quality of produced technology.

Keywords: MPS, chemistry, modeling, biology, engineering

1 CONTEXT OF OSE

1.1 Overview. As molecular modeling grows in impact and usage, the number of available tools also grows, but these often suffer from lack of quality or sustainability, with academic projects often ending at proof-of-concept stage without reaching the levels of maturity and robustness needed to benefit research and the economy. Such proof-of-concept work often suffers from poor documentation, inadequate or non-existent user support, and lack of long-term maintenance. Industry-based researchers face similar challenges with their customized, in-house software solutions, as they may not have the skills or time to update or extend the software. Proprietary software further prevents reuse and modification of existing solutions, and makes it harder to systematically benchmark and analyze simulation results, though such benchmarking is essential for systematic improvement of computational methods and models[1]. In summary, the further progress of molecular modeling requires robust and reliable research software regularly benchmarked for performance and accuracy. The open source approach to building common infrastructure is a great success story in the tech industry, and the molecular sciences will benefit from a similar model. Thus, we propose building better collaboration and coordination mechanisms to align researchers, developers and funders to effectively contribute to a research commons while fueling future discovery and innovation.

1.2 Guiding principles and long-term vision. The Open Molecular Software Foundation (OMSF) is a 501(c)(3) non-profit organization focused on facilitating collaboration and software development for molecular simulation and design. We believe that open source software and open science provide the best path to accelerated discovery and innovation, as permissive licenses enable rapid dissemination and democratize access to important tools and research results worldwide and across sectors. To ensure that our tools are built and maintained with quality, security and usability in mind, OMSF aims to facilitate collaboration between software teams, researchers, and funders to jointly design, build and validate key infrastructure for computational molecular sciences. Building OMSF as a community-driven organization is an integral part of our mission. Ultimately, we will create a strong and dynamic ecosystem of robust and widely-adopted open source infrastructure forming a “full stack” for molecular modeling, driving new discoveries and applications in molecular function and design. The full scope of our mission is given in Fig. 1.

1.3 Specific societal needs and broader impact. As the molecular sciences mature and key new innovations become critical for industry, we’re reaching a critical moment or turning point. By advancing open source innovation, we create a link between academic and industry researchers and software scientists which we will help translate into dramatic progress on our shared problems, ranging from materials design to food sciences and human health. The present moment is key, as new multi-billion dollar industries are launching, software and hardware tools are changing rapidly, and players are open to new, pre-competitive ways of approaching their shared problems.

1.4 Challenges. 1) Operational framework and coordination. Building communities requires a different set of skills than scientific software development, though both are needed to build sustainable open source infrastructure. Cross-institutional collaborations, in particular, require clear and solid foundations, including a clear legal framework around governance, management of funds and intellectual property rights, but also project and community management to assist with coordination of efforts across projects, institutions and individual contributors. The cost of coordination and efficient information exchange can be quite high, but failing to do so incurs even higher costs, resulting in missed collaboration opportunities, duplication of efforts and lack of interoperability. The latter has been a long standing challenge in the field[2]. OMSF recognizes the need for better coordination and management of people, resources, and assets associated with large, distributed infrastructure projects to achieve better outcomes for the entire ecosystem. For example, OpenFF recently led an unprecedented collaborative and standardized benchmarking effort of developed

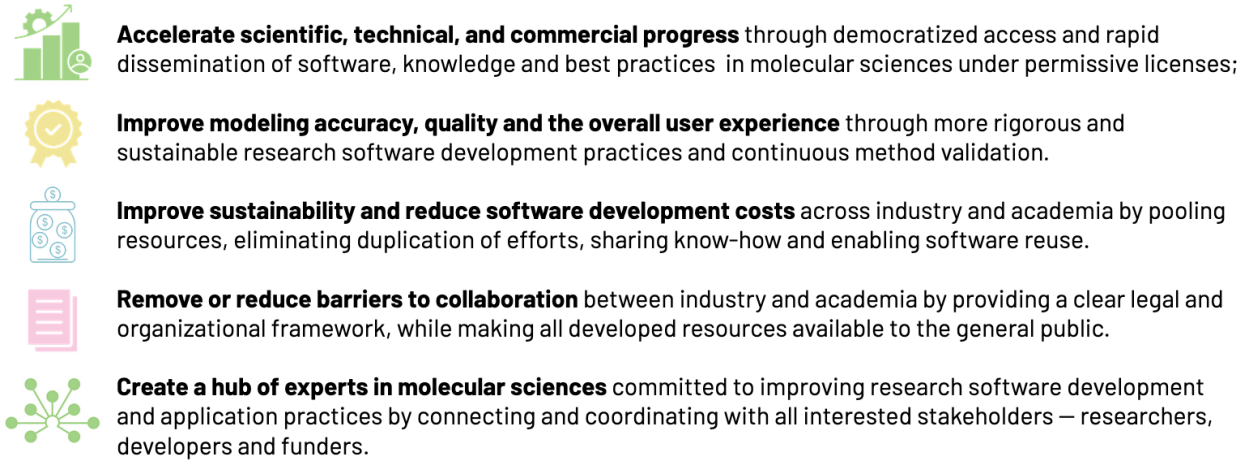

Figure 1: OMSF's mission and vision.

force fields on both public and proprietary datasets[3]. While the industry partners can't necessarily easily share information with each other, they can do so with us. We will leverage such neutral ground to enable cooperation across organizations. **2) Sustainability.** OMSF hosts several independent software projects, which have their own funding and governance, and OMSF takes a small fee to handle legal and administrative obligations (currently 10%). While OMSF's projects may already be sustainable, the organization itself is not; it is lean and has a minimal profile, yet because of this, it lacks resources to keep up with key demands from the community and to support its own growth. Specifically, we have no resources to invest into community building efforts and systematic improvements aimed at the entire ecosystem, though these are an integral part of our mission. Given OMSF's size and youth, it mostly operates in "firefighting" mode, dealing with urgent issues at the expense of investing in overall ecosystem health and sustainability. Thus currently, OMSF operates at a basic level with a single staff member supporting operational tasks, but we need additional investment to build community resources and ensure sustainability.

1.5 Aims. Via POSE, OMSF will build collaboration infrastructure to support open source projects in molecular sciences, and pave the way toward long-term ecosystem sustainability. In particular, we will develop several key infrastructure areas, including legal and contract infrastructure, contribution guidelines, best practices in software development and project governance, and easy access points for our prospective contributors, as well as streamlined onboarding. We need a one-time investment to create these materials and processes, which will yield high returns over time and a solid foundation for growth even after the project period. We expect long term benefits for the entire ecosystem. Sections 2-6 give a detailed overview of our aims and plans.

1.6 OMSF Ecosystem. In the roughly one year that OMSF has been in operation, it has helped nucleate two projects, and has been approached by several existing projects interested in pathways to sustainability, clearly demonstrating the need for OMSF. Table 1 lists active projects under the OMSF umbrella, along with projects in various stages of onboarding. All OMSF projects are required to release their software and data under OSI-approved permissive licenses for the broadest possible reuse. Typically, this means MIT for software and CC-BY 4.0 for data/know-how/other content. Collaboration and continuous assessment play a key role in project's progress with contributions from both industry and academic researchers[3;4].

Figure 2: Active OMSF projects (OpenFF[5–8], OpenFE[9;10], OpenFold[11;12]) and industry partners, and projects in different stages of onboarding (WESTPA[13–15], Rosetta[16;17], OpenMM[18;19], Fragalysis[20].

Active projects	**Industry partners**
Open Force Field (OpenFF), launched in Sep 2018, focuses on building accurate force fields (energy models) for use in molecular design, ranging from applications in the pharmaceutical industry to materials design. Its mission includes: 1) development of a modern, extensible, and sustainable framework for automated force field parameterization and application; 2) data management – generating, curating, and sharing datasets necessary for producing and benchmarking high-accuracy force fields; and 3) periodic releases of new comprehensive force fields. Force field accuracy is systematically improved through scientific innovation and use of large datasets. OpenFF already has a fairly extensive network of collaborators, interfacing with a science-oriented NIH-funded project led by Michael Shirts at Colorado and another open science project led by UKRI Future Leaders Fellow Danny Cole at Newcastle (UK), and its tools are in use in at least 15 pharmaceutical and software companies as well as by outside academic labs, such as that of Brian Space (NC State) in the materials design area.	**Current:** AbbVie, AstraZeneca, Bayer, Cresset, Eli Lilly, Janssen Pharmaceuticals, OpenEye Scientific Software, Pfizer, Redesign Science, Relay Tx, Roche, Ventus Tx, Vertex **Alumni:** BASF, Boehringer-Ingelheim Bristol Myers Squibb, Merck KGaA, GSK, Qulab, Xtalpi
Open Free Energy (OpenFE), launched in Sep 2021, develops tools for binding free energy calculations to guide pharmaceutical drug discovery and design. Its focus is on taking established academic science in the area of binding free energy techniques and bringing it to production level, where it can be used broadly, reducing redundancy and wasted effort. OpenFE allows the community to pool resources and build a single high quality set of tools with method benchmarking as a default part of every major release, while enabling direct feedback from researchers early in the development cycle.	AbbVie, AstraZeneca, Bayer, Boehringer Ingelheim, Bristol Myers Squibb, Confo Tx, Eli Lilly, Exscientia, Genentech, GSK, Interline Tx, Janssen, Merck KGaA Nurix Tx, Redesign Science, Relay Tx
OpenFold, launched in March 2022, seeks to develop a permissively licensed and community-governed AI technology competitive with the performance of state-of-the-art models, but with the entire training & inference stack provided under the same permissive license to enable model re-training and updates on commercially available GPUs. The early stage development focuses on development of single-sequence protein structure prediction algorithms inspired by AlphaFold.	**Current:** Arzeda, Bayer, Basecamp Research, Charm Therapeutics, Cyrus Biotechnology, Dassault Systemes, Outpace Bio, Prescient Design
Onboarding stage	**Status**
WESTPA is a high-performance Python framework for applying the weighted ensemble (WE) path sampling strategy, which enables simulations of processes that are orders of magnitude longer than the simulations themselves.	Governance model and membership agreement finalized.
Rosetta Commons. Rosetta Commons has just been awarded a POSE Phase I grant in support of their community transition to an OSE. We will collaborate closely with the Rosetta community in their exploration of transition pathways to open source licenses, project governance and sustainability models.	In process.
Fragalysis – a web-based platform for fragment-based drug discovery	In consideration
OpenMM – a high-performance toolkit for molecular simulation optimized for the latest generation of computer hardware, primarily GPUs	In consideration

2 ECOSYSTEM GROWTH

2.1. Building OSE through standardized templates and best practices. Every project comes with its own challenges and operating environments, but they often have common requirements for project tools and protocols. We will develop a set of recommendations and best practices around software development, project management, legal frameworks, and community building resources that every project will need, whether hosted by OMSF or not. Rather than addressing this for each individual project, we will develop a general guide for projects to follow, while also laying out how projects can effectively define their own "traffic system" and flow – an evolving **'open source playbook for the molecular sciences'** to enable the growth of our ecosystem while adhering to best practices around quality and security (3). We will further operationalize selected best practices and recommendations by developing a set of **customizable ecosystem templates** or blueprints, such as legal agreement templates, onboarding guides and covenants for contributors, and cookiecutters/kit designs for developers (see Sections 4 and 5). These templates and processes will create a mechanism for aligning people around decision-making processes and governance, and provide recommendations for communication and information dissemination strategies. Our ecosystem support team (Section 4) will drive development, implementation and regular assessment of developed resources during this grant period. We will disseminate the playbook content through public platforms and outreach activities. More importantly, we will establish direct feedback and contribution mechanisms to encourage reuse and active participation in future development and maintenance of these community resources across our ecosystem.

Figure 3: We plan a cyclical process where our community identifies certain processes and recommendations as best practices, and turns the most common ones into ecosystem templates.

2.2 Collaborative project development. Each of our current Hosted Projects already has a network of industry and academic researchers (see OpenFF, Open Free Energy, OpenFold websites) working together to build tools of interest together with the core developer teams. Our project teams usually involve a mix of OMSF-affiliated software scientists, academic and industry researchers. This allows us to efficiently explore new developments and rapidly incorporate expert feedback in every cycle of software development and validation, ensuring that we are developing products of value with an immediate positive impact on their research (see D.Cole's Collaboration Letter). The collaboration model that we currently employ usually involves the following steps: 1) collaborative roadmap planning; 2) new feature (capability) prototyping in academic groups or in collaboration with leading experts; 3) integration of new feature prototypes in the main codebase by core developer teams; 4) validation of new software features by academic and industry researchers; 5) maintenance of core packages by core developer teams[8]. Each of these steps provide an opportunity to contribute to the project and engage with the community members. We will combine these existing development processes with new resources developed in this grant, such as contributor and collaborator guides, plugin/kit designs, etc. to enable further community

engagement and growth of each project (see Section 4-6).

2.3 Technology and accessibility. We plan to make our infrastructure readily available and optimized for cloud deployment with the aim of making our tools more accessible to both academic and industry users via our partnership with Oracle Cloud (see Collaboration Letter). Oracle Cloud's robust infrastructure and the company's experience working with leaders in the research space will be a significant benefit towards OMSF's growth. A partnership with Oracle will provide a strong cloud framework that will continue to drive new capabilities and provide a means for researchers and software engineers, alike, to build novel and innovative open-source approaches in molecular sciences. This will also facilitate availability of OMSF software on other cloud providers.

2.4 Incentives and future growth potential. To date, OMSF projects have attracted interest from more than 30 pharmaceutical companies and startups, and recently from the tech industry. However, we are yet to develop a systematic approach for efficient recruiting of industry partners. A main motivation for industry researchers to support our projects is "scratching an itch", i.e. improving tools and their performance, pooling resources to reduce development cost and collaboratively assessing accuracy of methods and models on open and proprietary data. This accuracy assessment offers very important information for day-to-day research decisions, but comes at a high cost in terms of human and compute needed by any single group acting alone. It also helps with uncovering software bugs or any other performance issues that affect all stakeholders. These joint efforts build strong ties between industry researchers and trainees, resulting in more job opportunities for trainees and direct access to talent for industry. This was observed in practice within the Open Force Field Initiative, where several team alumni moved to positions with industry partners.

Academic science, too, benefits from stable and well-maintained infrastructure. With such tools, academics can focus on new science and discovery rather than building the necessary tooling, exploring new functionalities and research topics of interest instead. Trainee participation in OMSF projects opens up access to a network of experienced researchers in industry and academia, as well as the opportunity to closely collaborate with experienced research software engineers, benefiting all involved. Although we have dozens of academic researchers participating as collaborators, advisors and contributors across our current projects[?], we believe that this number can further grow by creating clear engagement pathways and self-onboarding mechanisms for researchers and contributors outside of our existing networks (see Section 4.1).

Our current approach to ecosystem growth relies heavily on our existing contact network and active outreach. Our projects are typically able to draw early adopters/testers by soliciting volunteers from these networks, which provide a solid foundation on which to build a sustainable ecosystem. We are also able to advertise new projects to our existing industry partners and facilitate collaboration or cost-sharing between projects when convenient. However, most of our existing projects are in the early stages and still need to build their communities of active users and contributors, thus we propose to expand and formalize our networks via POSE – it will allow us to use our existing resources and connections more efficiently and result in a long lasting impact on our sustainability. To illustrate, Micaela Matta (Kings College London) approached us with a funding opportunity and a desire to use OpenFF infrastructure for an exciting science project, but she needed an industry partner. We circulated her ideas to our industry network, and received interest from several partners. Ultimately she was able to partner with Janssen Pharmaceuticals to get a grant supporting the work of a Ph.D. student in this area, using OpenFF infrastructure to explore new science. Our collaboration with Danny Cole (Newcastle, see Collaboration Letter) was somewhat similar. We would like to see more of these examples where researchers pursue their own research interests while extending the OMSF ecosystem along the way. We will also aim to collaborate with our industry partners (e.g. Oracle for Research) and academic institutions (e.g MolSSI) on outreach by leveraging their existing networks to spur community growth and create

new connections.

3 ORGANIZATION AND GOVERNANCE

3.1 Current governance model. OMSF is a fully remote, distributed organization using fiscal sponsorship to engage in research software development activities and provide administrative and operational support for open source projects, with OMSF acting as a legal representative and custodian of participating codebases, and assisting with project governance, fundraising, distribution of funds, personnel, accounting, legal matters, etc. We seek to combine the **efficiency of centralized, shared organizational infrastructure with the benefits of the distributed, collaborative open source projects**. At the moment, our sponsored projects, called Hosted Projects, are the organization members. Each Hosted Project comes with its own membership agreement and governance structure. OMSF has a Board of Directors overseeing its activities, making strategic decisions and ensuring that the organization remains aligned with its tax-exempt purpose, with directors elected by the leadership of the hosted projects. Up to seven Directors may serve at the same time, where up to 4 "internal" members will represent the Hosted Projects and up to 3 "external" members who bring external expertise and represent the community. Two industry representatives will join the Board on Nov 1, 2022 and we expect further expansion in 2023 to reach the Board's maximum capacity. Our website[21] offers more details about the current Board members and OMSF governance.

The OMSF Board may approve new projects as Hosted Projects if they satisfy the following criteria: 1) Working on a problem within or related to the molecular sciences domain; 2) Licensing work under an OSI approved license for software, or under license compliant with Open Definition for other content; 3) Publicly available code and data on appropriate repositories (GithHub, GitLab, Zenodo, etc.); 4) Clear governance model and relationship with OMSF[21]. Once a project is approved, the project members sign an agreement specifying rules of participation and a charter describing the project's governance structure. OMSF can hire staff, give fellowships, apply for grants and accept donations on behalf of each Hosted Project. Most projects have a Governing Board making relevant decisions about project development, roadmaps, hiring, finances, etc. The OMSF Board of Directors provides oversight over Hosted Projects' activities to ensure alignment between projects and OMSF's missions, and adherence to legal and regulatory requirements (4A). This structure allows Hosted Project communities to manage and govern their own resources and responsibilities with admin and operational support provided by OMSF.

3.2 Membership model redesign. We will redesign OMSF's membership structure to 1) improve our governance model through better representation of our community, and 2) streamline our administrative processes. Our redesign will add a mechanism for formal participation in OMSF governance for meritorious contributors to ensure OMSF grows as a community-driven organization, where the community includes all contributors, not just those who contribute financially. We will also streamline the membership process for those joining multiple projects. At the moment, they must join each project separately, increasing the administrative burden on both sides (4 B,C). As we improve membership structure to address these limitations, we will also explore mechanisms to allow sponsors to divert a fraction of funding to key infrastructure which is critical for projects but which may not have its own support. OpenMM provides one example of such infrastructure (see Collaboration Letter)[18]. We will work together with our stakeholders to develop an acceptable membership model and implement the changes in our agreements. We will document and share our membership exploration process, along with our final plans and agreements, publicly.

3.3 General legal framework. In addition to the participation agreement for industry partners,

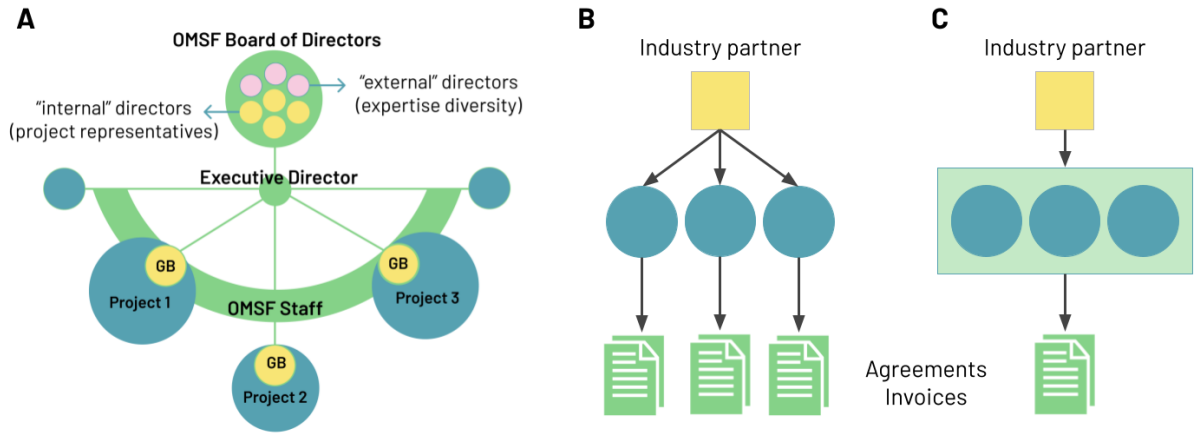

Figure 4: A) OMSF's current governance model; B) The current membership model requires a separate agreement and invoice for each project supported by the same industry partner, creating confusion on both ends; C) Proposed streamlining of the membership model by becoming an OMSF member and selecting projects of interest. This model also enables meritorious individuals to become OMSF members (omitted from the figure from clarity).

each Hosted Project requires a set of agreement templates to enable multi-institutional collaboration and clarify intellectual property (IP) rights and obligations. These include research collaboration agreements, fellowship agreements, MOUs, services agreements, contractor and contributor agreements, etc. We will make a one-time investment in a consistent legal framework with standardized terms and conditions around IP, and expected duties and responsibilities for frequently encountered scenarios. This approach will reduce our legal costs in the future by removing the need to negotiate similar agreements for every specific instance. Making resulting documents available for reuse can also reduce legal costs for other projects/organizations with similar needs, while providing transparency into our organizational operations. Finally, these standardized agreements will lead to faster setup times for new projects due to expedient legal reviews.

3.4 Operational streamlining. Following the membership model redesign, we will streamline the process of joining and renewing memberships for organizations and individuals alike, and that for donations. This will reduce our overall admin burden, improving OMSF's financial sustainability. We hope our efficient "admin stack" can become a valuable community resource, from better project support to collecting information needed for better understanding of our OSE economics.

4 COMMUNITY BUILDING

We will create clear access points for new community members, helping them to find their place in the ecosystem. The resulting materials and supporting activities will become vital infrastructure, serving OMSF and its projects in the long run and supporting healthy and sustainable projects. We will use POSE funding to build OMSF's community in several key areas listed below.

4.1 Engagement Pathways. Engagement Pathways are guides targeting areas in which we need to grow the OMSF community – bringing in new people to engage with and contribute to OMSF projects, bringing in new synergistic funding and science which can extend or help sustain our organization and projects, and helping new projects spring up or come on board and grow to become healthy and sustainable. We will deliver the following guides: **1) Contributor/Developer**

Guide – this pathway will focus on new software contributors, guiding them in how to get onboarded and begin contributing to OMSF projects (essentially, a developer guide), with an OMSF-level template being made available for projects to adapt; **2) Collaborator Guide** – this guide will be aimed at potential new collaborators interested in how OMSF projects/tools can be used to facilitate their science and how they can obtain OMSF help in facilitating their projects or obtaining their own funding. For example, OpenFF helped support Cole's successful proposal for a UKRI Future Leaders Fellowship (Collaboration Letter). This supports his group's work on independent science projects and their contributions back to OpenFF infrastructure and science, funding several full-time people in Cole's group. **3) Project Guide** – this guide will provide instructions for creating a new project or transferring an existing one under the OMSF umbrella, and show how OMSF can facilitate academic-industry partnerships focused on open science/open source software, while hosting projects benefiting the broad molecular science community.

4.2 Training, onboarding and growth pipelines. To grow our community, we will supplement contributor/developer guides with explicit training and mentoring programs. We will also work to design and set up industry internships to bring on new contributors, as many industry members have regular internship programs, but also through generating proposals to Google Summer of Code, Outreachy and similar efforts. Industry partners sometimes provide in-kind contributions to OMSF projects in terms of person-effort, further helping with dissemination of our tools and associated best practices. We will partner with other organizations and communities with comprehensive educational experience, such as Rosetta and MolSSI, to provide more training opportunities to new contributors. Our long-term goal is building a diverse,distributed and growing pipeline of researchers who use, value, and contribute to our tools and projects, whether or not they maintain any official connection with OMSF (see also Section 2.3).

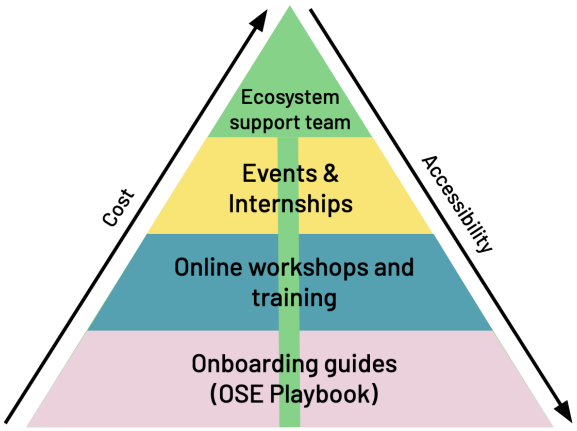

Figure 5: We will build our community through self-onboarding by making our contributor/project guides and training materials available broadly, and by actively building those resources and offering training through our ecosystem support team for maximum impact.

4.3 Outreach/Events. OMSF and our projects use websites, social media platforms, and software development platforms to communicate with our users and contributors. Our projects also organize regular online workshops to provide an overview of the recent progress and future plans, and software demos. We plan to organize at least one in-person event per year to provide an opportunity for our fully remote organization, project stakeholders and adjacent communities to spend some time together, form new relationships, and strengthen the existing ones. We will use these events to showcase our projects, but also to tackle common challenges such as best practices, infrastructure gaps, project management, contributor training, and sustainability.

4.4 Ecosystem support team. We will establish a dedicated ecosystem support team consisting of an ecosystem manager and research software engineers ("ecosystem engineers") whose primary responsibility will be development and maintenance of shared resources important for the entire OSE ("ecosystem commons", see Section 5). They will also assist with integration of developed tools and practices across community projects, providing support and training when needed. We expect that this process will

result in wider adoption of best practices and improved flow of ideas and expertise across the ecosystem. Ultimately, we believe that this approach will lead to improved ecosystem interoperability, systematic upskilling of project contributors, and easier cross-project contributions. On the outward-facing side, we will assist our projects with market research to understand their audience, as well as identifying potential synergies and evaluating costs/benefits associated with those synergies, and guide them to effectively communicate with their community. This will be the primary duty of the ecosystem manager. We will improve our ecosystem and expand individual projects' capacity through a combination of resource development and active support by the dedicated team.

5 OSE DEVELOPER TOOLS (ECOSYSTEM INFRASTRUCTURE)

We will also build several key developer tools to address certain security and quality control challenges at the ecosystem level (see below). All our Hosted Projects strive to follow OpenSSF Best Practices whenever possible (see Data Management Plan)– using GitHub for version control and management of development, pull request reviews for merging new features; comprehensive documentation, code review, continuous integration and automated testing, nightly builds, roadmap planning, etc. However, these practices require constant attention to ensure that everything is working as intended as different dependencies are being updated, and contributors need training and guidance as they come with diverse skills and backgrounds. We will design and build a set of tools and processes aiming to improve and/or standardize certain development practices and contribution requirements, as a part of our contributor onboarding materials and developer guides. Our goal is to facilitate compliance with OpenSSF Best Practices through technical solutions and training available to our projects and broader community. To ensure future relevance and sustainability of the tools built in this project, we will create a community governance model around each product so it can "exit to community".

5.1 Reusable plugin/kit framework. To improve extensibility, we will build a framework to allow projects to easily accept plugin code. An existing "primary" package can be enhanced in two different ways without having to modify the original source code: Plugin code dynamically hooks into an existing code base (typically using a registry and callback mechanism) and enhances or changes the functionality. Alternatively, a kit is a self-contained computational tool or formal workflow description that relies on the primary package. The main difference between plugins and kits is the depth of integration, with kits being more loosely coupled than plugins. Plugins or kits are separate from the development of the main package and remain under the control of the original authors. The separation allows for a different pace of development and it enhances the recognition of the plugin/kit authors. The major advantage for the development team of the primary package is that they do not shoulder the burden of maintenance and they avoid accumulation of technical debt, thus freeing them to focus on core functionality of the main package. This division of responsibility is essential for scaling up the growth of an ecosystem around a package because otherwise the number of developers of the main package would have to be increased to keep pace with the addition of new features or risk the degradation of software quality, leading to unmaintainable and unsustainable code.

Although anyone can write plugins/kits, they often fail to attract users unless they are visible and findable, accessible, installable via standard mechanisms, and continuously tested (following FAIR software standards[22]). As achieving these standards is difficult for small development teams and scientists with little or no formal training in software development, it is essential that plugin/kit developers are supported in order to grow a plugins/kits ecosystem. This support can come in the form of a framework that is customized for the primary package. The two major com-

ponents of such frameworks are (1) code templates and (2) a registry mechanism. Examples for such package-specific frameworks are PLUMEDNest[23] (input files for enhanced sampling simulations with PLUMED[24]); the Aiida plugin registry (workflows in material science)[25], the napari-hub[26] (plugins to enhance the napari image viewer[27]), and MDAKits[28] (enhancements of MDAnalysis[29] via kits and plugins). Different projects can implement plugins/kits in different ways. However, common approaches and common problems exist. As part of this grant we will provide a general foundational framework that addresses key problems for plugins/kits and will make it easy for a project to bootstrap its own plugin/kit ecosystem.

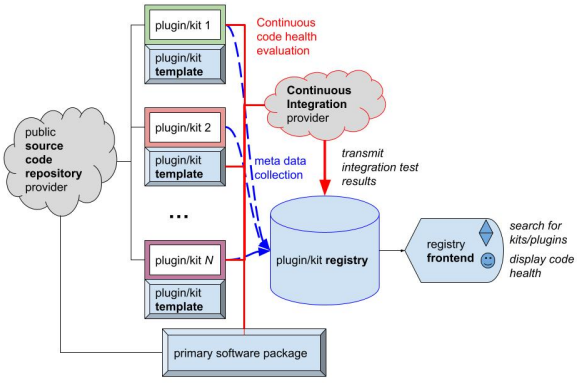

Figure 6: Generic plugin/kit ecosystem. Plugins/kits associated with a primary software package are registered in a registry. Integration between primary package and plugins/kits is continuously tested with continuous integration workflows whose results are stored in the registry. The registry frontend allows users to search for plugins/kits and view the current code health.

The basic components of our generic framework (6) are: 0) template documentation for requirements that plugins/kits must fulfill, 1) a template for a plugin/kit package that encodes best practices [see also Improve, extend and build resources for quality assurance below], 2) a registry backend for maintaining a record of kits/plugins, 3) continuous integration workflows for running integration tests between plugin/kit and the primary package, 4) a registry frontend that displays available plugins/kits together with appropriate code health metrics (e.g., version compatibility, test results, test coverage) and allows users to find appropriate plugins/kits. All of these components will be made available so that they can be easily customized and run within the organization of the primary package (e.g., their own GitHub Organization). We will leverage a common core of tools that is used by OMSF projects and most of the biomolecular software community: Python is used either as the primary programming language or as a glue language. Projects perform distributed version control via git and GitHub and use continuous integration (via GitHub actions) for testing and in some cases deployment.

5.2 Distribution systems/Packaging. Our projects only matter when they are in the hands of users and researchers. As such, a solid distribution system is required to deliver software products and retrieve dependencies. Additional benefits of using these distribution systems come in the form of improved interoperability and findability of related tools, support for a range of architectures and automation/inheritance of certain best practices and dev strategies. Enabling more tools to join the same distribution platform(s) will further strengthen our ecosystem. We already deploy to platforms such as the Python Package Index? (PyPI), and conda-forge[30] (CF), but packaging new releases can take up substantial developer time due lack of expertise or platform-related issues. CF is of special interest here, but our developers report that CF documentation can be confusing, that some key CF packages sometimes remain broken for extended periods of time, and that pulling back a broken package can be challenging. We will make deployment easier for our teams and the broader community through training, improved documentation, bug fixing and new feature implementation, including contributing back to the CF community. To increase the value and impact of these updates, we will first survey open source communities and developers in our ecosystem

to find the most pressing issues associated with CF. We will also survey our industry partners and other users to better understand how we could improve their experience. We will share the results of this survey and work on fixing the most important problems identified by the community. We will work with Quansight (see Collaboration Letter) and NumFocus consultants to deliver solutions and training, and make CF expertise more accessible. This cooperation will lead to improved software quality, security, and availability across the entire ecosystem.

5.3 Continuous integration/Automated testing. In addition to distribution systems, one of the common challenges for OMSF projects is the ability to test software on diverse hardware and operating systems. Access to diverse hardware is paramount to adequate testing, yet there is not a reliable service to test on hardware like GPUs, Apple Silicon (ARM64 architecture or M1 chips), or other less common architecture. We will run Self-Hosted Runners of GitHub actions as all the codebases are on GitHub. These runners will handle automatic continuous integration (CI) testing on production-level hardware, unavailable normally. The objective is not simply to have a single or limited set of hardware to run OMSF projects, but to serve as proof-of-concept for infrastructure we can make available to the broader community once successful enough as containerized workflows that users can deploy on hardware of their choice. We will partner with Oracle Cloud to enable access to different hardware architectures. These resources, and importantly how to construct and maintain them, will persist far beyond our own projects or the timeframe of the award.

5.4 Cookiecutter update. We will collect and enshrine key proposed changes inside an updated version of the Molecular Sciences Software Institute's (MolSSI) "Cookiecutter for Computational Molecular Sciences" (CC-CMS) package[?], as well as improve its maintenance. At its core, the CC-CMS is a molecular-sciences-centric tool for building Python packages from scratch which automatically sets up best practices tools and ideas such as CI, automatic documentation, semantic versioning, dependency resolution, testing frameworks, and more. Each of the above mentioned improvements and extensions can, and will, be incorporated into the CC-CMS as they help large swaths of the scientific community. The tools and configurations for establishing Self-Hosted runners on arbitrary hardware will be built into the CI setup. Updated features that we commision for CF will be included in the deployment files, and we'll streamline the deployment process to go from Python library to deployment on CF. This work will be done in collaboration with Levi Naden and MolSSI (see Collaboration Letter).

6 SUSTAINABILITY

Our goal is to scale and support our **ecosystem as a distributed network connected via shared infrastructure and best practices**, something like a "traffic system" for our OSE, by aligning development practices. We will leverage templates and processes to create "self-onboarding" routes and empower different communities to establish new projects, adhere to best practices in software development and governance, actively contribute to community resources, and drive innovation[31]. This will facilitate collaboration and broaden the contributor pool, since participants will be using the same tooling and developer guides (Section 3-5) making engagement far easier. We need dedicated personnel to create and implement these templates and processes, as well as pathways for engaging with the broader community – in terms of sharing and implementing templates and processes developed in this grant, attracting and onboarding new contributors, bringing in new projects and funders, and creating synergistic projects which can attract their own funding (an area of prior success). We will build a system which can operate long after the award ends.

For successful delivery and future relevance, we will continuously improve our processes based on deliberate project scoping and specification, and clear communication and engagement with

relevant stakeholders at every step of the life cycle. We will assign a project driver (or lead team) to each proposed deliverable to ensure that our goals are met on time. More information about our development and evaluations plans, including success metrics, are provided below.

6.1 OSE Playbook I: Legal framework, governance and operations. With a one-time investment of time and effort, we will create a framework to enable fast and easy deployment of new collaborative projects with clear, adaptable governance models; mechanisms to specify and/or formalize relationships and responsibilities between participating institutions or individual contributors; and an inclusive membership and governance model for OMSF (Section 3). We will also develop an auditing process and ensure license compliance for all our software products relying on external open source components to avoid licensing issues for our downstream users. We will further improve our operational efficiency through streamlined administrative processes.

Success metrics: We will measure the impact of provided templates and services via surveys, community engagement (web traffic and social media), number of downloads and references made to shared materials, number of projects adopting our legal templates and governance models, and improved administrative capacity, i.e. larger volume of work supported by the same amount of effort.

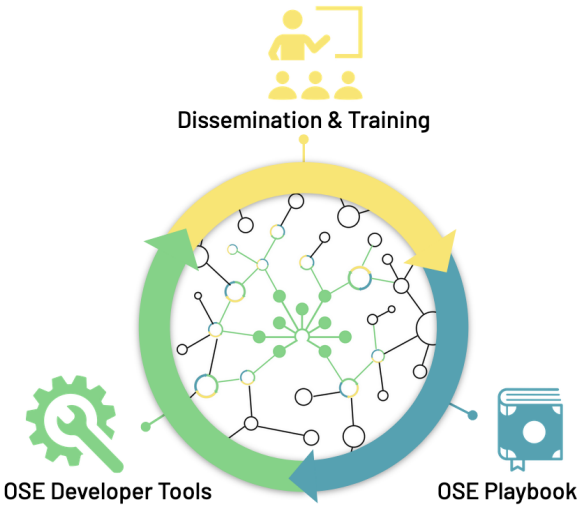

Figure 7: The concept of sustainability for our ecosystem is based on driving continuous improvement and adoption of best practices, and inclusion and community growth through collaboration and training.

Evaluation plan. 1) Legal framework. We will evaluate our new membership model and legal framework via surveys and interactive meetings with our industry partners and academic collaborators to ensure our agreements fulfill their intended purpose in a form acceptable to our partnering institutions – a clear sign of success. This means that applicable new agreements will have to be reviewed by legal teams of our industry partners and academic institutions as a part of the development process. We will run systematic reviews at least once a year during our annual membership renewals to assess and improve our legal framework and updated released documents accordingly. **2) Operational streamlining.** Our administrative team meets bi-weekly to assess and review the current workload and supporting processes in a relatively complex landscape of procurement systems employed by our industry partners. As we improve our stack and processes, we will monitor administrative efficiency as a success metric.

6.2 OSE Playbook II: Onboarding guides. We will develop three onboarding guides (Section 4.1) – for contributors/developers, collaborators (researchers) and projects. The latter will also include more generalized project management recommendations based on OpenFF protocols and processes. We will develop these resources as a living document on GitHub, enabling version control and direct contributions from our community in the form of issues or pull requests, following the same principles as "enhancement proposals"[??]. We will gather feedback and lessons learned to create major releases of these guides at least once a year. The ecosystem manager will lead development of these resources.

Success metrics: Community engagement and adoption (web traffic, social media, pull re-

quests, issues, forks, etc), number and diversity of contributors, downloads/citations.

Evaluation plan. Before releasing the initial version of the onboarding guides, we will request review from our project members and close collaborators. After the release, we will accept feedback on GitHub and allow modifications by external contributors. We will assign a lead maintainer to each guide and require quarterly review of guides.

6.3 OSE developer tools. As described in Section 5, we will: 1) design and develop a reusable plugin/kit framework; 2) improve automated testing for different hardware platforms; 3) make *conda-forge* distribution system more accessible; and 4) update cookiecutter for easy setup of new projects with best practices embedded. This infrastructure will help with systematic improvement of best practices and their dissemination/implementation, resulting in an overall improvement of software quality and security for the entire ecosystem. We will hire a team of research software engineers ("ecosystem engineers") to lead and actively support development, maintenance and integration of these tools for the benefit of the entire ecosystem, as well as assist with training and knowledge transfer to further expand software development skills in current and future contributors. They will receive assistance from our project teams, collaborators (O. Beckstein, L. Naden, J. Rodriguez-Guerra) and consultants when needed, as well as from the ecosystem manager and PIs.

Success metrics: Success metrics for all software components will include general open source project health metrics, including number of downloads/forks/issues/pull requests[32], number of contributors, number of dependencies, publication citations, imports of the software into external projects (e.g. "from openff.toolkit import x"), user satisfaction assessment, workshop/event attendance, number of views/downloads for training materials, and requests for support.

Adoption and evaluation. Best practices change with time and technology, requiring our approach to change. We take a two-pronged approach to development and adoption: **1) Working groups** to design and review software and/or best practices. We will establish a working group around each item listed in Section 5 to enable early community engagement and a path to long-term sustainability by defining clear roles and responsibilities in future development and maintenance. The working groups will assist with roadmap planning and specification of deliverables, provide feedback on design and implementation, and meet regularly to discuss progress and best practices (at least quarterly). These groups will also examine technology releases, updated software, and potential future changes which will happen between meetings, and recruit contributors and future members. Ecosystem engineers will be responsible for early development and deployment of this infrastructure, as well as providing long-term maintenance. With this approach, we aim to jumpstart a community of developers and maintainers dedicated to improvement of OSE developer tools and best practices, and through that, quality and security of open source software in molecular modeling. **2) Deployment of new changes.** Deploying changes to ecosystem infrastructure will require several feature advancements that we will develop. First, an upgrade path will need to be planned to automatically apply potential changes, especially because the broader user communities will have wildly divergent codes at the time of upgrades. The upgrade paths are similar to table upgrades in SQL databases. Wherever possible, we will develop a bot to automatically deploy these changes to existing repositories. The working groups will also document each upgrade path designed for human consumption so users can easily adopt newer best practices when automatic update is not possible.

6.4 Community building. While we hope that the developed infrastructure under 6.1 and 6.2 will enable self-onboarding for many new contributors, we will engage in community building (Section 4). We will **promote the developed guides, tools and practices** through social media, and workshops and demos aimed at the broader benefits of the above resources. We will work with our partnering institutions and communities (MolSSI, QuanSight, Oracle, Rosetta, industry partners) to **host workshops for different skill level users and create internship opportunities** in part-

nership with our industry partners and through existing programs, such as GSoC. Our ecosystem engineers will spend up to 20% on training and upskilling research engineers in our ecosystem. We will further work to expand our existing networks through new collaborations, synergistic funding opportunities, and joint efforts around common infrastructure. Our annual in-person events will allow us to make stronger personal connections with our community members, but also to reduce friction inherent to remote work in a globally distributed environment. These efforts will be led by the ecosystem support team with assistance from the PIs, senior personnel and other collaborators.

Success metrics: a) New contributors and collaborators joining OMSF communities; b) Increasing or stable number of total open source projects under OMSF's umbrella; c) Increasing number of new organizations supporting OMSF/Hosted Projects; d) Size and diversity of OMSF team to support the ecosystem, including diversity of roles, expertise and backgrounds; f) Amount of funding raised annually for OMSF and hosted project activities; g) Amount of synergistic funding OMSF helped to raise; h) Number of industry/GSoC internships to build trainee pipeline; i) Number of online and in-person events and workshops to train new users and contributors, attendance levels; j) Creation of useful training materials combined with available user engagement metrics (web traffic, comments, social media), k) Web traffic metrics, number of active forum users and active conversation threads, followers on social media and engagement levels with our content.

Evaluation plan. While each specific metric listed above may be frequently evaluated and updated, we will gather all available information and run a systematic review of our community building strategy once a year. This process will require significant effort and involvement of multiple stakeholders. Each Hosted Project will also track their own success metrics and work on their own outreach and community building strategies, and report back to the ecosystem support team at least once a year. Success of events and workshops will be evaluated after the event and findings will be recorded for the annual review. Based on the insights gained through suggested metrics and annual review, we will update our strategy and reallocate our effort as needed.

7 BROADER IMPACTS

The community recognizes a vital need for quality open source software to sustain scientific innovation, and is ready to invest money and effort in this space when shown a way forward, as our current hosted projects have demonstrated. This has driven the success of OpenFF and OpenFE projects, and inspired OpenFold, but we at OMSF see a huge potential to do more and better not just for our projects, but for the entire computational molecular sciences ecosystem – we want to catalyze a culture change around collaboration and resource sharing in molecular sciences. Here, we request the resources to and make our tools and know-how available and understandable to anyone who would like to build on or extend the OMSF model. This grant will allow OMSF, and the broader community, to speed up development of essential support systems for the open source ecosystem in molecular sciences to drive future growth and systematic improvements to software, best practices, and contributor onboarding and training. The value and promise of future successes of computational modeling in molecular sciences is only rising, as demonstrated by the recent successes in protein structure prediction by AlphaFold2 and other ML applications and the recent biotech investment boom. Our work will create opportunities for (under-resourced) researchers to make an impact on this exciting technology by linking them with funding, talent and contributions. Our work will also create new roles and career paths for scientists who love open source and open science, acting as a bridge between industry and academia. The ultimate outcome of this POSE project will be a sustainable organization, well integrated with the community, which helps nucleate and sustain a broad range of open source projects in the molecular sciences,

and welcomes all contributors regardless of career stage, geographic location, or community.

References

[1] Walters, W. P. Code Sharing in the Open Science Era. *Journal of Chemical Information and Modeling*, 60 (10):4417, 2020. https://doi.org/10.1021/acs.jcim.0c01000, 10.1021/acs.jcim.0c01000, publisher: American Chemical Society.

[2] Condic-Jurkic, K. Molecular Dynamics Software Interoperability Workshop Report. *Technical report*, Zenodo, 2021. https://zenodo.org/record/7152685, 10.5281/zenodo.7152685.

[3] D'Amore, L. and Hahn, D. Follow-up workshop on Benchmarking, 2021. https://zenodo.org/record/5369858, 10.5281/zenodo.5369858.

[4] Hahn, D., Bayly, C., Boby, M. L., Macdonald, H. B., Chodera, J., Gapsys, V., Mey, A., Mobley, D., Benito, L. P., Schindler, C., Tresadern, G., and Warren, G. Best Practices for Constructing, Preparing, and Evaluating Protein-Ligand Binding Affinity Benchmarks [Article v1.0]. *Living Journal of Computational Molecular Science*, 4 (1):1497, 2022. https://livecomsjournal.org/index.php/livecoms/article/view/v4i1e1497, 10.33011/livecoms.4.1.1497, number: 1.

[5] Qiu, Y., Smith, D. G. A., Boothroyd, S., Jang, H., Hahn, D. F., Wagner, J., Bannan, C. C., Gokey, T., Lim, V. T., Stern, C. D., Rizzi, A., Tjanaka, B., Tresadern, G., Lucas, X., Shirts, M. R., Gilson, M. K., Chodera, J. D., Bayly, C. I., Mobley, D. L., and Wang, L.-P. Development and Benchmarking of Open Force Field v1.0.0—the Parsley Small-Molecule Force Field. *Journal of Chemical Theory and Computation*, 17 (10):6262, 2021. https://doi.org/10.1021/acs.jctc.1c00571, 10.1021/acs.jctc.1c00571, publisher: American Chemical Society.

[6] Open Force Field software repository. https://github.com/openforcefield.

[7] Open Force Field website. https://openforcefield.org.

[8] Open Force Field knowledge base. https://openforcefield.atlassian.net/wiki/spaces/FF/overview.

[9] Open Free Energy software repository. https://github.com/OpenFreeEnergy.

[10] Open Free Energy website. https://openfree.energy.

[11] Ahdritz, G., Bouatta, N., Kadyan, S., Xia, Q., Gerecke, W., and AlQuraishi, M. OpenFold, 2021. https://zenodo.org/record/5709539, 10.5281/ZENODO.5709539.

[12] OpenFold website. https://openfold.io.

[13] Russo, J. D., Zhang, S., Leung, J. M. G., Bogetti, A. T., Thompson, J. P., DeGrave, A. J., Torrillo, P. A., Pratt, A. J., Wong, K. F., Xia, J., Copperman, J., Adelman, J. L., Zwier, M. C., LeBard, D. N., Zuckerman, D. M., and Chong, L. T. WESTPA 2.0: High-Performance Upgrades for Weighted Ensemble Simulations and Analysis of Longer-Timescale Applications. *Journal of Chemical Theory and Computation*, 18 (2):638, 2022. https://doi.org/10.1021/acs.jctc.1c01154, 10.1021/acs.jctc.1c01154, publisher: American Chemical Society.

[14] Zwier, M. C., Adelman, J. L., Kaus, J. W., Pratt, A. J., Wong, K. F., Rego, N. B., Suárez, E., Lettieri, S., Wang, D. W., Grabe, M., Zuckerman, D. M., and Chong, L. T. WESTPA:

an interoperable, highly scalable software package for weighted ensemble simulation and analysis. *Journal of Chemical Theory and Computation*, 11 (2):800, 2015. 10.1021/ct5010615.

[15] westpa. https://github.com/westpa.

[16] RosettaCommons software repository. https://github.com/RosettaCommons.

[17] RosettaCommons website. https://www.rosettacommons.org/.

[18] Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., and Pande, V. S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Computational Biology*, 13 (7):e1005659, 2017. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005659, 10.1371/journal.pcbi.1005659, publisher: Public Library of Science.

[19] OpenMM. OpenMM software repository. https://github.com/openmm.

[20] Fragalysis, G. xchem. https://github.com/xchem.

[21] Open Molecular Software Foundation website, 2022. https://omsf.io.

[22] Chue Hong, N. P., Katz, D. S., Barker, M., Lamprecht, A.-L., Martinez, C., Psomopoulos, F. E., Harrow, J., Castro, L. J., Gruenpeter, M., Martinez, P. A., and Honeyman, T. FAIR Principles for Research Software (FAIR4RS Principles). *Research Data Alliance*, 2021. https://zenodo.org/record/6623556#.YqCJTJNBwlw, 10.15497/RDA00068, publisher: Research Data Alliance.

[23] Bonomi, M., Bussi, G., Camilloni, C., Tribello, G. A., Banáš, P., Barducci, A., Bernetti, M., Bolhuis, P. G., Bottaro, S., Branduardi, D., Capelli, R., Carloni, P., Ceriotti, M., Cesari, A., Chen, H., Chen, W., Colizzi, F., De, S., De La Pierre, M., Donadio, D., Drobot, V., Ensing, B., Ferguson, A. L., Filizola, M., Fraser, J. S., Fu, H., Gasparotto, P., Gervasio, F. L., Giberti, F., Gil-Ley, A., Giorgino, T., Heller, G. T., Hocky, G. M., Iannuzzi, M., Invernizzi, M., Jelfs, K. E., Jussupow, A., Kirilin, E., Laio, A., Limongelli, V., Lindorff-Larsen, K., Löhr, T., Marinelli, F., Martin-Samos, L., Masetti, M., Meyer, R., Michaelides, A., Molteni, C., Morishita, T., Nava, M., Paissoni, C., Papaleo, E., Parrinello, M., Pfaendtner, J., Piaggi, P., Piccini, G., Pietropaolo, A., Pietrucci, F., Pipolo, S., Provasi, D., Quigley, D., Raiteri, P., Raniolo, S., Rydzewski, J., Salvalaglio, M., Sosso, G. C., Spiwok, V., Šponer, J., Swenson, D. W. H., Tiwary, P., Valsson, O., Vendruscolo, M., Voth, G. A., White, A., and The PLUMED consortium. Promoting transparency and reproducibility in enhanced molecular simulations. *Nature Methods*, 16 (8):670, 2019. https://www.nature.com/articles/s41592-019-0506-8, 10.1038/s41592-019-0506-8, number: 8 Publisher: Nature Publishing Group.

[24] Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C., and Bussi, G. PLUMED 2: New feathers for an old bird. *Computer Physics Communications*, 185 (2):604, 2014. https://www.sciencedirect.com/science/article/pii/S0010465513003196, 10.1016/j.cpc.2013.09.018.

[25] Huber, S. P., Zoupanos, S., Uhrin, M., Talirz, L., Kahle, L., Häuselmann, R., Gresch, D., Müller, T., Yakutovich, A. V., Andersen, C. W., Ramirez, F. F., Adorf, C. S., Gargiulo, F., Kumbhar, S., Passaro, E., Johnston, C., Merkys, A., Cepellotti, A., Mounet, N., Marzari, N., Kozinsky, B., and Pizzi, G. AiiDA 1.0, a scalable computational infrastructure for automated reproducible

workflows and data provenance. *Scientific Data*, 7 (1):300, 2020. https://www.nature.com/articles/s41597-020-00638-4, 10.1038/s41597-020-00638-4, number: 1 Publisher: Nature Publishing Group.

[26] Chan Zuckerberg Initiative. napari hub. https://www.napari-hub.org/about.

[27] Sofroniew, N., Lambert, T., Evans, K., Nunez-Iglesias, J., Bokota, G., Winston, P., Peña-Castellanos, G., Yamauchi, K., Bussonnier, M., Doncila Pop, D., Can Solak, A., Liu, Z., Wadhwa, P., Burt, A., Buckley, G., Sweet, A., Migas, L., Hilsenstein, V., Gaifas, L., Bragantini, J., Rodríguez-Guerra, J., Muñoz, H., Freeman, J., Boone, P., Lowe, A., Gohlke, C., Royer, L., PIERRÉ, A., Har-Gil, H., and McGovern, A. napari: a multi-dimensional image viewer for Python, 2022. https://doi.org/10.5281/zenodo.3555620, 10.5281/zenodo.3555620.

[28] Alibay, I., Barnoud, J., Beckstein, O., Gowers, R. J., Naughton, F., and Wang, L. MDAKits: Supporting and promoting the development of community packages leveraging the MDAnalysis library, 2022. https://figshare.com/articles/preprint/MDAKits_Supporting_and_promoting_the_development_of_community_packages_leveraging_the_MDAnalysis_library/20520726/1, 10.6084/m9.figshare.20520726.v1.

[29] Gowers, R. J., Linke, M., Barnoud, J., T. J. E. Reddy, Melo, M. N., Seyler, S. L., Dotson, D. L., Domanski, J., Buchoux, S., Kenney, I. M., and Beckstein, O. MDAnalysis: A Python package for the rapid analysis of molecular dynamics simulations. In S. Benthall and S. Rostrup, editors, *Proceedings of the 15th Python in Science Conference*, pages 102–109. Austin, TX, 2016. http://conference.scipy.org/proceedings/scipy2016/oliver_beckstein.html, 10.25080/Majora-629e541a-00e.

[30] Community, C.-F. The conda-forge Project: Community-based Software Distribution Built on the conda Package Format and Ecosystem. 2015. https://zenodo.org/record/4774216, 10.5281/ZENODO.4774216, publisher: Zenodo.

[31] Leman, J. K., Weitzner, B. D., Renfrew, P. D., Lewis, S. M., Moretti, R., Watkins, A. M., Mulligan, V. K., Lyskov, S., Adolf-Bryfogle, J., Labonte, J. W., Krys, J., Consortium, R., Bystroff, C., Schief, W., Gront, D., Schueler-Furman, O., Baker, D., Bradley, P., Dunbrack, R., Kortemme, T., Leaver-Fay, A., Strauss, C. E. M., Meiler, J., Kuhlman, B., Gray, J. J., and Bonneau, R. Better together: Elements of successful scientific software development in a distributed collaborative community. *PLOS Computational Biology*, 16 (5):e1007507, 2020. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1007507, 10.1371/journal.pcbi.1007507, publisher: Public Library of Science.

[32] Gardner, P. P., Paterson, J. M., McGimpsey, S., Ashari-Ghomi, F., Umu, S. U., Pawlik, A., Gavryushkin, A., and Black, M. A. Sustained software development, not number of citations or journal choice, is indicative of accurate bioinformatic software. *Genome Biology*, 23 (1):56, 2022. https://doi.org/10.1186/s13059-022-02625-x, 10.1186/s13059-022-02625-x.

[33] Duarte Ramos Matos, G. and Mobley, D. L. Challenges in the use of atomistic simulations to predict solubilities of drug-like molecules. *F1000Research*, 7:686, 2018. https://f1000research.com/articles/7-686/v1, 10.12688/f1000research.14960.1.

[34] Liu, S., Cao, S., Hoang, K., Young, K. L., Paluch, A. S., and Mobley, D. L. Using MD Simulations To Calculate How Solvents Modulate Solubility. *Journal of Chemical Theory and Computation*, 12 (4):1930, 2016. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=26878198&retmode=ref&cmd=prlinks, 10.1021/acs.jctc.5b00934.

[35] Bannan, C. C., Calabró, G., Kyu, D. Y., and Mobley, D. L. Calculating Partition Coefficients of Small Molecules in Octanol/Water and Cyclohexane/Water. *Journal of Chemical Theory and Computation*, 12 (8):4015, 2016. https://pubs.acs.org/doi/10.1021/acs.jctc.6b00449, 10.1021/acs.jctc.6b00449.

[36] Yang, M., Dybeck, E., Sun, G., Peng, C., Samas, B., Burger, V. M., Zeng, Q., Jin, Y., Bellucci, M. A., Liu, Y., Zhang, P., Ma, J., Jiang, Y. A., Hancock, B. C., Wen, S., and Wood, G. P. F. Prediction of the Relative Free Energies of Drug Polymorphs above Zero Kelvin. *Crystal Growth & Design*, 20 (8):5211, 2020. https://doi.org/10.1021/acs.cgd.0c00422, 10.1021/acs.cgd.0c00422, publisher: American Chemical Society.

[37] Klimovich, P. V., Shirts, M. R., and Mobley, D. L. Guidelines for the analysis of free energy calculations. *Journal of Computer-Aided Molecular Design*, 29 (5):397, 2015. https://link.springer.com/article/10.1007/s10822-015-9840-9, 10.1007/s10822-015-9840-9.

[38] Klimovich, P. V. and Mobley, D. L. A Python tool to set up relative free energy calculations in GROMACS. *Journal of Computer-Aided Molecular Design*, 29 (11):1007, 2015. https://doi.org/10.1007/s10822-015-9873-0, 10.1007/s10822-015-9873-0.

[39] Bannan, C. C., Burley, K. H., Chiu, M., Shirts, M. R., Gilson, M. K., and Mobley, D. L. Blind prediction of cyclohexane–water distribution coefficients from the SAMPL5 challenge. *Journal of Computer-Aided Molecular Design*, 30 (11):927, 2016. http://link.springer.com/10.1007/s10822-016-9954-8, 10.1007/s10822-016-9954-8, tex.ids: Bannan:2016:J.Comput.AidedMol.Des.

[40] Rustenburg, A. S., Dancer, J., Lin, B., Feng, J. A., Ortwine, D. F., Mobley, D. L., and Chodera, J. D. Measuring experimental cyclohexane-water distribution coefficients for the SAMPL5 challenge. *Journal of Computer-Aided Molecular Design*, 30 (11):945, 2016. http://link.springer.com/10.1007/s10822-016-9971-7, 10.1007/s10822-016-9971-7.

[41] Zanette, C., Bannan, C. C., Bayly, C. I., Fass, J., Gilson, M. K., Shirts, M. R., Chodera, J. D., and Mobley, D. L. Toward Learned Chemical Perception of Force Field Typing Rules. *Journal of Chemical Theory and Computation*, 15 (1):402, 2019. https://doi.org/10.1021/acs.jctc.8b00821, 10.1021/acs.jctc.8b00821.

[42] Mobley, D. L., Bannan, C. C., Rizzi, A., Bayly, C. I., Chodera, J. D., Lim, V. T., Lim, N. M., Beauchamp, K. A., Slochower, D. R., Shirts, M. R., Gilson, M. K., and Eastman, P. K. Escaping Atom Types in Force Fields Using Direct Chemical Perception. *Journal of Chemical Theory and Computation*, 2018. https://doi.org/10.1021/acs.jctc.8b00640, 10.1021/acs.jctc.8b00640.

[43] Işık, M., Levorse, D., Rustenburg, A. S., Ndukwe, I. E., Wang, H., Wang, X., Reibarkh, M., Martin, G. E., Makarov, A. A., Mobley, D. L., Rhodes, T., and Chodera, J. D. pKa measurements for the SAMPL6 prediction challenge for a set of kinase inhibitor-like fragments. *Journal of Computer-Aided Molecular Design*, 32 (10):1117, 2018. http://link.springer.com/10.1007/s10822-018-0168-0, 10.1007/s10822-018-0168-0, tex.ids: Isik:2018:J.Comput.AidedMol.Des.a.

[44] Loeffler, H. H., Bosisio, S., Duarte Ramos Matos, G., Suh, D., Roux, B., Mobley, D. L., and Michel, J. Reproducibility of Free Energy Calculations across Different Molecular Simulation Software Packages. *Journal of Chemical Theory and Computation*, 14 (11):5567, 2018. https://doi.org/10.1021/acs.jctc.8b00544, 10.1021/acs.jctc.8b00544.

[45] Mobley, D. L. and Gilson, M. K. Predicting Binding Free Energies: Frontiers and Benchmarks. *Annual Review of Biophysics*, 46 (1):531, 2017. http://www.annualreviews.org/doi/10.1146/annurev-biophys-070816-033654, 10.1146/annurev-biophys-070816-033654.

[46] Shirts, M. R., Klein, C., Swails, J. M., Yin, J., Gilson, M. K., Mobley, D. L., Case, D. A., and Zhong, E. D. Lessons learned from comparing molecular dynamics engines on the SAMPL5 dataset. *Journal of Computer-Aided Molecular Design*, 31 (1):147, 2017. https://link.springer.com/article/10.1007/s10822-016-9977-1, 10.1007/s10822-016-9977-1.

[47] Yin, J., Henriksen, N. M., Slochower, D. R., Shirts, M. R., Chiu, M. W., Mobley, D. L., and Gilson, M. K. Overview of the SAMPL5 host–guest challenge: Are we doing better? *Journal of Computer-Aided Molecular Design*, 31 (1):1, 2017. http://link.springer.com/article/10.1007/s10822-016-9974-4, 10.1007/s10822-016-9974-4.

[48] Paluch, A. S., Parameswaran, S., Liu, S., Kolavennu, A., and Mobley, D. L. Predicting the excess solubility of acetanilide, acetaminophen, phenacetin, benzocaine, and caffeine in binary water/ethanol mixtures via molecular simulation. *The Journal of Chemical Physics*, 142 (4):044508, 2015. http://scitation.aip.org/content/aip/journal/jcp/142/4/10.1063/1.4906491, 10.1063/1.4906491.

[49] Rizzi, A., Murkli, S., McNeill, J. N., Yao, W., Sullivan, M., Gilson, M. K., Chiu, M. W., Isaacs, L., Gibb, B. C., Mobley, D. L., and Chodera, J. D. Overview of the SAMPL6 host–guest binding affinity prediction challenge. *Journal of Computer-Aided Molecular Design*, 32 (10):937, 2018. https://doi.org/10.1007/s10822-018-0170-6, 10.1007/s10822-018-0170-6.

[50] Gill, S. C., Lim, N. M., Grinaway, P. B., Rustenburg, A. S., Fass, J., Ross, G. A., Chodera, J. D., and Mobley, D. L. Binding Modes of Ligands Using Enhanced Sampling (BLUES): Rapid Decorrelation of Ligand Binding Modes via Nonequilibrium Candidate Monte Carlo. *The Journal of Physical Chemistry B*, 122 (21):5579, 2018. https://pubs.acs.org/doi/10.1021/acs.jpcb.7b11820, 10.1021/acs.jpcb.7b11820.

[51] Riquelme, M., Lara, A., Mobley, D. L., Verstraelen, T., Matamala, A. R., and Vöhringer-Martinez, E. Hydration Free Energies in the FreeSolv Database Calculated with Polarized Iterative Hirshfeld Charges. *Journal of Chemical Information and Modeling*, 2018. https://doi.org/10.1021/acs.jcim.8b00180, 10.1021/acs.jcim.8b00180.

[52] Fox, G., Qiu, J., Crandall, D., von Laszewski, G., Beckstein, O., Paden, J., Paraskevakos, I., Jha, S., Wang, F., Marathe, M., Vullikanti, A., and Cheatham, Thomas E., I. Contributions to High-Performance Big Data Computing. In L. Grandinetti, G. R. Joubert, K. Michielsen, S. L. Mirtaheri, M. Taufer, and R. Yokota, editors, *Future Trends of HPC in a Disruptive Scenario*, volume 34 of *Advances in Parallel Computing*, pages 34–81. IOS Press, 2019.

[53] Fox, G., Glazier, J. A., Kadupitiya, J. C. S., Jadhao, V., Kim, M., Qiu, J., Sluka, J. P., Somogyi, E., Marathe, M., Adiga, A., Chen, J., Beckstein, O., and Jha, S. Learning Everywhere: Pervasive Machine Learning for Effective High-Performance Computation. In *2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 422–429. 2019. 10.1109/IPDPSW.2019.00081.

[54] Khoshlessan, M., Paraskevakos, I., Jha, S., and Beckstein, O. Parallel Analysis in MDAnalysis using the Dask Parallel Computing Library. In K. Huff, D. Lippa, D. Niederhut,

and M. Pacer, editors, *Proceedings of the 16th Python in Science Conference*, pages 64–72. Austin, TX, 2017. http://conference.scipy.org/proceedings/scipy2017/mahzad_khoslessan.html, 10.25080/shinma-7f4c6e7-00a.

[55] Fan, S., Linke, M., Paraskevakos, I., Gowers, R. J., Gecht, M., and Beckstein, O. PMDA - Parallel Molecular Dynamics Analysis. In C. Calloway, D. Lippa, D. Niederhut, and D. Shupe, editors, *Proceedings of the 18th Python in Science Conference*, pages 134 – 142. Austin, TX, 2019. https://conference.scipy.org/proceedings/scipy2019/shujie_fan.html, 10.25080/Majora-7ddc1dd1-013.

[56] Paraskevakos, I., Luckow, A., Khoshlessan, M., Chantzialexiou, G., Cheatham, T. E., Beckstein, O., Fox, G., and Jha, S. Task-parallel Analysis of Molecular Dynamics Trajectories. In *ICPP 2018: 47th International Conference on Parallel Processing, August 13–16, 2018, Eugene, OR, USA*, page Article No. 49. ACM, New York, NY, USA, 2018. 10.1145/3225058.3225128, backup Publisher: Association for Computing Machinery.

[57] Khoshlessan, M., Paraskevakos, I., Fox, G. C., Jha, S., and Beckstein, O. Parallel performance of molecular dynamics trajectory analysis. *Concurrency and Computation: Practice and Experience*, 32:e5789, 2020. 10.1002/cpe.5789.

[58] Jakupovic, E. and Beckstein, O. MPI-parallel Molecular Dynamics Trajectory Analysis with the H5MD Format in the MDAnalysis Python Package. In M. Agarwal, C. Calloway, D. Niederhut, and D. Shupe, editors, *Proceedings of the 20th Python in Science Conference*, pages 40–48. Austin, TX, 2021. https://conference.scipy.org/proceedings/scipy2021/edis_jakupovic.html, 10.25080/majora-1b6fd038-005.

[59] Ma, H., Bhowmik, D., Lee, H., Turilli, M., Young, M., Jha, S., and Ramanathan, A. Deep Generative Model Driven Protein Folding Simulations. *Parallel Computing: Technology Trends*, pages 45–55, 2020. https://ebooks.iospress.nl/doi/10.3233/APC200023, 10.3233/APC200023, publisher: IOS Press.

[60] Casalino, L., Dommer, A. C., Gaieb, Z., Barros, E. P., Sztain, T., Ahn, S.-H., Trifan, A., Brace, A., Bogetti, A. T., Clyde, A., Ma, H., Lee, H., Turilli, M., Khalid, S., Chong, L. T., Simmerling, C., Hardy, D. J., Maia, J. D., Phillips, J. C., Kurth, T., Stern, A. C., Huang, L., McCalpin, J. D., Tatineni, M., Gibbs, T., Stone, J. E., Jha, S., Ramanathan, A., and Amaro, R. E. AI-driven multiscale simulations illuminate mechanisms of SARS-CoV-2 spike dynamics. *The International Journal of High Performance Computing Applications*, 35 (5):432, 2021.

[61] Dotson, D. L., Seyler, S. L., Linke, M., Gowers, R. J., and Beckstein, O. datreant: persistent, Pythonic trees for heterogeneous data. In S. Benthall and S. Rostrup, editors, *Proceedings of the 15th Python in Science Conference*, pages 51 – 56. Austin, TX, 2016. http://datreant.org, 10.25080/Majora-629e541a-007.

[62] Kenney, I. M., Beckstein, O., and Iorga, B. I. Prediction of cyclohexane-water distribution coefficients for the SAMPL5 data set using molecular dynamics simulations with the OPLS-AA force field. *J Comput Aided Mol Des*, 30 (11):1045, 2016.

[63] Beckstein, O., Fox, G., and Jha, S. Convergence of data generation and analysis in the biomolecular simulation community. In *Online Resource for Big Data and Extreme-Scale Computing Workshop*, page 4. 2018. https://www.exascale.org/bdec/sites/www.exascale.org.bdec/files/whitepapers/Beckstein-Fox-Jha_BDEC2_WP_0.pdf.