Foldit Drug Design Game Usability Study: Comparison of Citizen and Expert Scientists

Yunchao Liu
Elect. Engr. & Comp. Science
Vanderbilt University
yunchao.liu@vanderbilt.edu

Rocco Moretti
Center for Structural Biology
Department of Chemistry
Vanderbilt University
rmorettiase@gmail.com

Bobby Bodenheimer
Elect. Engr. & Comp. Science
Vanderbilt University
bobby.bodenheimer@vanderbilt.edu

Jens Meiler
Dept. of Biochemistry
Vanderbilt University
jens.meiler@vanderbilt.edu

ABSTRACT

In building a new drug design mode for the popular citizen scientist game Foldit, we focus on creating an easy-to-use and intuitive interface to confer complex scientific concepts to citizen scientist players. We hypothesize that to be efficient in the hands of citizen scientists such an interface will look different from well-established drug-design software used by experts. We used the relaxed think-aloud method to compare citizen and expert scientists working with our prototype interface for Foldit Drug Design Mode (FDDM). First, we tested if the two groups are providing different feedback when it comes to the usability of the prototype interface. Second, we investigated how the difference between the two groups might inform a new game design. As expected, the results confirm that experienced scientists differ from citizen scientists in engaging their background knowledge when interacting with the game. We then provided a prioritization list of background knowledge employed by the expert scientists to derive design suggestions for FDDM.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in interaction design**.

KEYWORDS

usability, citizen science, game design, drug discovery

ACM Reference Format:
Yunchao Liu, Rocco Moretti, Bobby Bodenheimer, and Jens Meiler. 2020. Foldit Drug Design Game Usability Study: Comparison of Citizen and Expert Scientists. In *Motion, Interaction and Games (MIG '20), October 16–18, 2020, Virtual Event, SC, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3424636.3426899

MIG '20, October 16–18, 2020, Virtual Event, SC, USA

ACM ISBN 978-1-4503-8171-0/20/10…$15.00
https://doi.org/10.1145/3424636.3426899

1 INTRODUCTION

Citizen science projects engage the public in scientific research problems that benefit from extensive human parallelization. Citizen science projects have been applied to a broad range of topics such as biology, archaeology, and astronomy [Silvertown 2009]. Volunteers around the globe contribute to science by collecting species data they observe in the wild [Sullivan et al. 2009], by donating their idle computing power for analyzing data [Korpela et al. 2001; Shirts and Pande 2000], by labeling data with tags [Crowston and Prestopnik 2013], and so on.

In this paper we focus on the gamification of citizen science projects. Such games fall into a sub-class of serious games known as games with a purpose (GWAP) [von Ahn 2006; Von Ahn and Dabbish 2008], or human computation games (HCGs) [Siu et al. 2017]. The gamification brings in benefits such as socialization [Iacovides et al. 2013], competition [Bowser et al. 2013], and enjoyment [Von Ahn and Dabbish 2008] to the traditional citizen science projects. The usability of these games is essential to attract more potential citizen scientists and to get high-quality data [Bowser et al. 2013].

We use the online multiplayer citizen science game Foldit[1] for our study. Initially designed for protein folding [Cooper et al. 2010], Foldit has enabled citizen scientists to come up with protein structure prediction algorithms comparable to their expert counterparts' discovery [Khatib et al. 2011a] and has expanded its scope to several protein-related scientific problems, including de novo protein design [Koepnick et al. 2019], protein prediction with electron density maps [Horowitz et al. 2016], and comparative modeling [Khatib et al. 2011b].

The present study is part of an effort to build a new drug design mode of Foldit, termed "Foldit Drug Design Mode (FDDM)." Drug discovery is a lengthy and costly process. Companies spend over 10 years to bring a new drug to market from initial hit-to-lead optimization, lead-to-drug development, clinical trials to approval [DiMasi et al. 2010]. This is in part due to time-consuming search in the vast chemical space that a small-molecule drug can reside. We hypothesize that the creativity of citizen scientists can be harnessed to find the starting points for new drugs and thus to expedite the early stages of drug discovery.

[1] https://fold.it/

FDDM aims to create a platform for the player to design a small molecule compound that could potentially be developed into a therapeutic after optimization and validation in the later development stages. This new mode involves both a small-molecule (the potential drug being designed) and a protein (the target of which the biological function is to be inhibited or activated). While most previous Foldit puzzles focus on engineering proteins, Foldit players have successfully remodeled a Diels-Alderase to increase its activity [Eiben et al. 2012]. Since correctly predicting the interaction of the enzyme with its small-molecule substrate was a necessary component of this work, this success demonstrates that Foldit players are not only able to work on proteins alone, but also to work with small-molecule/protein interfaces providing an initial proof of principle for our FDDM approach.

For many scientific problems targeted by citizen science games, alternative software tools designed for professional scientists exist. This is particularly true for Computer-Aided Drug Design (CADD) because of its importance in research and discovery in academia and industry. Examples include AutoDock [Morris et al. 2009], GOLD [Verdonk et al. 2003], SeeSAR.[2], and even Rosetta, the biochemical engine which is used by Foldit [Meiler and Baker 2006]. However, we hypothesize that the most efficient interface for a citizen scientist will look different from an interface designed for an expert scientist. While we want to emulate the functionalities of expert CADD interfaces, we need to display these functionalities in an intuitive way to be effectively leveraged by citizen scientists. Thus we seek to answer two research questions with this study:

RQ1 What differences exist in the comparative usability needs of expert and citizen scientists?

RQ2 How does the comparison inspire a new design flow for citizen science games?

We used the relaxed-think aloud method [Ericsson and Simon 1984] to examine the answers to the above questions. Our results show that citizen and expert scientists' feedback differ mainly in the background knowledge expressed during the usability study. The results also suggest that the modes panel, the action panel, and the view panel within FDDM need to be made more user-friendly.

2 BACKGROUND

With the help of digital devices, thousands of citizen science projects around the globe are engaging millions of individuals in science [Bonney et al. 2014]. Gamification is a powerful tool to provide citizen science projects with enhanced user experience and engagement. Foldit was created to harness this human three-dimensional spatial reasoning to predict protein structure [Cooper et al. 2010]. Foldit players successfully came up with the accurate model of the Mason-Pfizer monkey virus retroviral protease [Khatib et al. 2011b]. In this paper, we are using Foldit in the context of small molecule drug discovery. Small molecule therapeutics are low molecular weight compounds (<1kDa) that have a potent biological effect [Stanczyk et al. 2008].

This paper studies the interface for a small molecule drug discovery game from the perspective of experts and citizen scientists. Experts have the training to typically judge whether a compound is promising candidate for further validation. Meanwhile, citizen scientists can often think out of the box and are excellent at distributed tasks. These observations lead to the question: can we design a citizen science game that combines the strength of both player groups? Recent work has shown that these groups view problems in different ways [Miller et al. 2019] but that both groups have strengths [Heck et al. 2018]. While knowledge gaps between expert scientists and novice citizens may make it unlikely for a system to satisfy both user groups [Crowston and Prestopnik 2013], we believe it is important to try and to understand the differences in a citizen science game.

3 METHOD

To evaluate the user interface, we use a think-aloud technique [Jaspers 2009; Jaspers et al. 2004], conceptually diagrammed in Figure S1.

3.1 Participants

We use opportunity sampling [Jupp 2006] to recruit participants. The citizen scientist group was recruited at our university from general chemistry classes, organic chemistry classes, and rotation students in our laboratory. The expert scientist group was recruited from those with practical drug design experience in our laboratory and other collaborators both at our university and at partner institutions. Neither groups have experience with the FDDM as it is not publicly available. Ethical approval was obtained from the Institutional Review Board of our university. Participants reviewed and signed a consent form before the study began.

Twenty-two participants took part in this study, 12 in the citizen scientist group and 10 in the expert scientist group. Sixteen participants were male, six female. The ages of participants ranged from 18 to 34 years old for the novice group (M = 22.3, SD = 5.3), and from 23 to 47 years old for the expert group (M = 29.7, SD = 7.4).

3.2 Game Play

The puzzle used for the study is 1573b HIV protease Inhibitor Small Molecule Design: Round 3b.[3] This puzzle presents the player with a protein target (HIV protease), and a starting small molecule, also termed the ligand or inhibitor (see Figure 1). The player's task is to modify the ligand both chemically and geometrically to achieve as high a score as possible within 30 minutes. The target backbone cannot be modified, but the target sidechains near the ligand can be moved. Rosetta [Leaver-Fay et al. 2011] calculates the score in the backend based on the chemical nature of the ligand and its interactions with the target protein. While the energy computed by Rosetta is negative-better [Cooper et al. 2010], Foldit inverts and rescales the energy to match the positive-better nature of most other game scores. A high score suggests that this small molecule might bind to the protein and might have an effect in inhibiting or altering the target protein function.

To maximize the score, the player would need to optimize the Rosetta-predicted interactions of the ligand with the target protein. Often this may take the form of positioning in certain functional groups or fragments with known good interactions with a particular environment. However, maximizing interaction energy is not sufficient, as the game also includes bonuses for compounds in the reasonable range of physicochemical properties such as solubility,

[2]https://www.biosolveit.de/SeeSAR/

[3]https://fold.it/portal/node/2006171

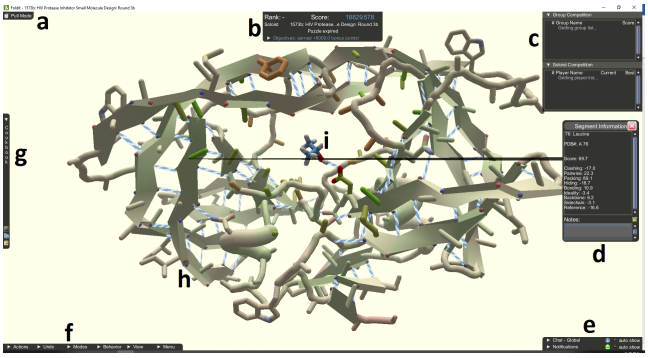

Figure 1: A screenshot of the game interface showing overall layout of the interface for the puzzle used in the test. (a) current mode state; (b) score panel; (c) leaderboard panel; (d) sidechain panel, only visible when the player hovers the cursor over the sidechain and presses 'tab'; (e) chat panel; (f) control bar, which allows access to actions panels, undo panel, modes panel, behavior panel, view panel and menu panel; (g) cookbook panel, not available in this puzzle; (h) target protein; (i) ligand, in the protein's binding pocket. The goal of the puzzle is to change the ligand to optimally fit the binding pocket of the surrounding protein.

number of rotatable bonds, and molecular weight, which are known to be important in practice for drugs, but are poorly considered by the raw Rosetta score.

Within the game, the player seeks to maximize the score by modifying the ligand to improve the ligand-protein interaction through a set of possible actions. These include adding or deleting atoms/fragments/bonds, changing the conformations of the protein sidechains near the pocket or of the ligand. 'Shake' allows optimized change of the configuration of the ligand to get a higher score; 'wiggle' allows optimized change of its position to get a higher score. 'Pull mode' allows manual change of the conformation or position, where the score may increase or decrease. Additionally, the ligand can be translated, historical best scores can be recalled, and various view options can be turned on/off. The game also contains other tools not relevant to drug design.

3.3 Procedure

The study was conducted either in person in our lab or remotely where communication was managed via Zoom. At the start of the study, the experimenter introduced the goal of the game, the tasks for the session, how the think-aloud process worked, provided a description of the HIV protease inhibitor puzzle, and educated the player on what actions they have in the game. Participants were welcome to ask questions during both consent form reading and introductory sessions.

During the gameplay, the participant practiced think-aloud, describing their thoughts as they played the game. The experimenter primarily listened, but gave hints in the following scenarios:

(1) The player explicitly asks for a specific function that exists in the game. (On the initial request, the experimenter would encourage exploration. In the case of repeated requests, the experimenter would provide hints as to the function's location.)

(2) The player does not use essential tools, without which they are unlikely to make a substantial progress. Examples of these essential tools were the 'undo' functionality, the small molecule design palette, and the 'wiggle' local energy optimization. The undo panel lets the player go back to previous moves. The small molecule design palette allows the player to add atoms/bonds/fragments to the ligand manually. The wiggle tool uses the minimization function in Rosetta to automatically increase the score for a given structure designed by the player.

(3) Certain known bugs.

(4) Players stray away from the main task, e.g., attempt to exit the puzzle or pursue protein folding tools instead of the drug design ones.

4 ANALYSIS & RESULTS

To answer RQ1,[4] we first examined the video recordings from all study participants and identified 793 comments. Identified comments exclude verbalizations such as reading from the screen, filler words (e.g. "um"), and exclamations (e.g. "Oh!"). Then, based on the comment contents , we created actionable categories and labeled the comments accordingly. "Actionable" is intended to reflect that each category should correspond to potential actions in future refactoring efforts. Table S1 shows examples from each category.

The 'confusion' category contains comments that show the player does not know how to use a function or does not know what in-game information means. For this category, we can provide more precise and complete information/guidance in the redesign.

The 'background knowledge' category contains comments that express knowledge not given in the game. For this category, we can provide useful knowledge explicitly in the game to guide players in the redesign.

The 'want' category contains comments that show the player's desire to have functionalities or information that are currently not available in the game. For this category, we might develop a new tool or provide the information needed.

The 'like/dislike' category contains comments that express a liked or disliked aspect of the game. For this category, we can promote liked features and discourage or alter disliked features in the refactoring.

The 'hint' category contains comments that contain experimenter hints. Traditionally, the analysis of think-aloud does not include what the experimenter says. However, the prompts provided by the experimenter to the player in the relaxed think-aloud method provides valuable information about which hints are useful in-context. These hints can form the basis for in-game prompts.

The 'bug' category contains comments that are related to bugs. For this category, we need to fix the mentioned bugs.

The 'inference' category contains comments that express knowledge not directly present in the game but can be inferred from in-game resources. For this category, we can give more direct guidance or information representation in the refactoring.

[4]RQ1: In a relaxed think-aloud usability study, what is the difference between an expert and citizen scientist feedback?

The 'frustration' category contains incidents where the player gives up on trying a functionality, typically accompanied by an unsatisfied tone. For this category, we can address the issues leading to the frustration.

Table S3 shows the total number of comments for each group (citizen, expert) by category. We ran a two-sample t-test to compare the mean difference between these totals for the groups in each category. The results of this analysis (p-values, mean difference between groups, and 95% confidence intervals) are also shown in Table S3. Raw data and Gardner-Altman Plots can be found in supplementary material in Table S2 and Figures S3-S10.

Only the 'background knowledge' category has a statistically significant p-value, i.e., expert scientists have a significantly higher mean in background knowledge than citizen scientists. For other categories, we fail to see such a statistically significant difference.

To answer RQ2,[5] we looked at the two sub-questions:

Sub-question1 What aspect of the game were mentioned most frequently by players?

Sub-question2 What can we do to improve these aspects?

To prepare the dataset for analysis, we aggregate similar comments to create unique descriptions (UDs). For example, "Still not sure what this color stands for. Does red stand for oxygen?" and "I wish I could see which atom is which" both express the players' confusion on the atom identity, and thus these two comments were assigned the same UD, "Confusion on the atom identity". This process results in 338 UDs. We then discarded UDs with comments from only one or two participants, since we decided there was too much randomness for individual comments to consider directly. This thresholding resulted in 56 UDs comprised of 440 total comments (see Figure S2). The UDs thus allow us to hierarchically categorize the comments without dealing with the vagaries of how these ideas were expressed.

To answer sub-question1, we examined comments and found the topics of comments either involve a chemistry concept, or a game element. We termed the first one the 'chemistry concept' group and the second one the 'game element' group. The 'chemistry concept' group has 181 total comments, and 'game element' group has 259 comments.

For the 'chemistry concept' group, four topics were identified: 'general chemistry', 'drug design', 'protein', and 'score'. We defined a comment to be in the 'general chemistry' groups when it concerned with atoms, functional groups, or bonds (single /double /triple /hydrogen). We defined a comment to be in the 'drug design' group if it is related to drug physicochemical properties, common motifs, geometrical and space considerations, or ligand-target interactions. The 'protein' groups contained comments that are about sidechains, secondary structures, and other directly protein-related topics. The 'score' groups was for comments relevant to energy functions.

As shown in Figure 2, 'general chemistry'(94 comments) and 'drug design'(68 comments) account for 89.5% of the total comments in the 'chemistry concept' group, and thus we focus on these two subgroups. Table S4 shows the distribution of comments among 'general chemistry' and 'drug design' subgroups.

[5]RQ2: How does the comparison inspire a new design flow for citizen science games?

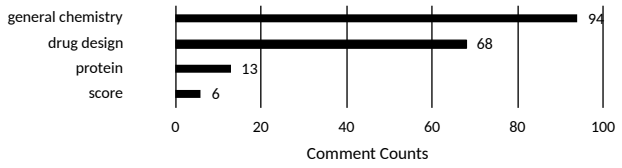

Figure 2: Number of comments in the 'chemistry concept' group.

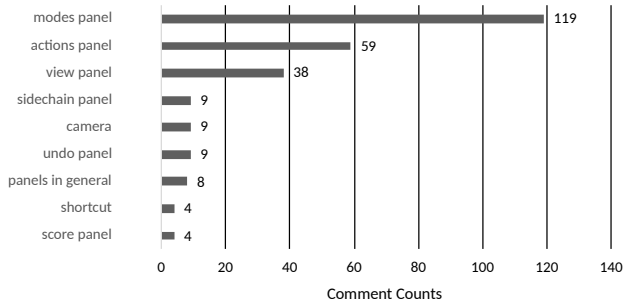

Figure 3: Number of comments in the 'game element' group.

In the 'general chemistry' subgroup, 'atoms' received the most comments (46 comments), comprising 49% of this subgroup (94 comments). The 'hydrogen bond' follows and makes up 36% (34 comments) of the subgroup total. In the 'drug design' subgroup, 'space' is preeminent with 80.9% (55 comments) of the total subgroup(68 comments).

In the 'game element' group (see Figure 3), we found that the 'modes panel' subgroup has the most comments (119 comments), more than twice that of the second largest subgroup. Also, the three highest count subgroups account for 83.4% (216 comments) of the group total (259 comments). To gain a better understanding of these groups, Figures S11-S14 in supplementary material show screenshots of these panels, and Table S5 shows the individual components of these three subgroups.

Examining the 'modes panel' subgroup (119 comments), 90% (107 comments) concern 'ligand design', and 10% (12 comments) concern the 'pull mode', which allows manual change of the position or conformation of the ligand. Since clicking the ligand design icon opens the small molecule design palette immediately, the small molecule design palette is the main topic. In the 'actions panel' subgroup (59 comments), 70% (41 comments) concern the 'wiggle' functionality, and 31% (18 comments) concern the 'shake' functionality. The 'view panel' subgroup(38 comments) has more diversified topics. The 'show clashes' accounts for the most significant component, which constitutes 37% (14 comments) of the subgroup total.

To improve the aspects of the game that concerned the players most (sub-question2), we examine those comments in categories that are actionable, i.e., that represent comments upon which the game can be changed. From Table S4, the dominant group in the 'background knowledge' category is the 'chemistry concept' group,

which is not surprising as most concepts could be expressed as background knowledge. Expert scientists also tend to express more background knowledge than citizen scientists, and this corresponds to our analysis with RQ1. Some concepts are expressed in confusion. In contrast to more background knowledge expressed by expert scientists, citizen scientists have more comments that reflect their confusion than the expert. Comparing 'atoms' and 'hydrogen bond' subgroups, we note that there is more confusion expressed in the 'atoms' subgroup, while there is more background knowledge expressed in the 'hydrogen bond' subgroup. No confusion was expressed in either the 'fragment' or 'bond' groups. Overall, then, we can see that the 'confusion' and the 'background knowledge' are the most prominent categories, as they comprise about 2/3 of the total comments. This fact is visualized by a sunburst chart in Figure S2 in supplementary material.

Consistent with RQ1, the difference between the expert and citizen scientists in 'game element' group is not conspicuous. This observation is consistent with the participant pool in that neither experts nor citizen scientists had prior experience with the FDDM interface itself. Moreover, there were no comments in the 'background knowledge' category, the 'want' category, or in the 'dislike' subcategory. For the 'background knowledge' category, it is most likely to be expressed as a concept group type, which we analyzed above. We also noticed that there is a total of 54 comments in the 'want' category before we apply the "discard comments that are from less than three participants" filter, and 15 left after the filter. The filtering rate is 72%, which means most desired functions and information have only been expressed by one or two participants. Similar to the 'dislike' subcategories, the filtering rate is 100%, indicating that all negative comments came from less than three participants. This finding could be a result of noise or just a low frequency event. In the 'ligand design' of the 'modes panel' subgroup, we see that the difference between citizen and expert scientists is not salient. We note that both citizen and expert scientists are confused and frustrated by these FDDM functionalities in the prototype. For the 'wiggle' of the 'actions' panel, citizen scientists tend to have received many more hints. Citizen scientists expressed more confusion and inference for the 'show clashes' function in the 'view panel' subgroup as well.

5 DISCUSSION

It is not surprising that background knowledge is the main difference between citizen scientists and expert scientists, given the greater experience and formal education of the latter. However, as both citizen and expert scientists see similar rates of the other categories of expression, this difference might indicate issues with the interface, rather than scientific concepts, which drives the current deficiencies of the game. The high prevalence of expressions in the "confusion" category might suggest that more explicit guidance and easier-to-use tools are needed. It is also worth noticing that there is no agreement in the comments about what is disliked ('dislike' subcategories have zero counts after filtering). While this may point to participants liking or at least being ambivalent to the interface (despite their confusion), it may also indicate that participants are less willing to make negative comments under the current study design.

The details of the comments provide insight into improving the game. Here we focus on the design palette as an example, as it has a large proportion of the UD counts. If we look at each comment in this subgroup, we recognize that both citizen and expert scientists are confused or frustrated in using the palette to modify the ligand.

Currently, the way to modify a ligand is as follows: (1) click on the ligand to select the atom(s) you want to modify (Figure S15 in supplementary material). (2) After selection, the elements and fragments on the design panel will light up, and if more than one atom is selected, bonds on the design panel will light up while fragments will gray out (Figure S16). (3) When the mouse hovers over the fragment, a preview of it will show on the ligand. (Figure S17) (4) Left click on the atom to replace the currently selected atoms on the ligand. Left click on a fragment will add it to the selected atom. Sometimes an addition is not applicable due to chemical restriction. (5) To delete, an atom needs to be selected by left clicking, and then the delete button needs to be clicked. However, the added fragment can only be deleted atom by atom, not as a whole. It is not possible to delete hydrogens by design. The reason why a fragment cannot be deleted once added is that there is no way to tell it apart from the rest of the ligand after addition. Hydrogen atoms are space holders to fulfill the bonding requirement of the connected atom, and thus cannot be deleted.

By analyzing the comments, we thus identified that the main reasons for people finding the design palette hard to use are: (1) not knowing they should click the ligand first to select it — Without the selection, the element and fragment icons on the palette are grayed out and not clickable; (2) being unsure how to delete either because of being unable to find the delete button, mistaking clearing selection icon for removing (see Figure S14 for the delete button), or thinking the delete button could remove the whole fragment; (3) not knowing chemical restrictions, with the error prompt only saying the addition is chemically infeasible, instead of specifying a precise reason; (4) selecting only one atom, so bonds on the design panel are grayed out and not clickable; and, (5) trying to delete hydrogens but failing.

For other aspects of the game, we suggest design improvements to provide more information about how to use tools, and make available more background knowledge as identified above. We also learned what parts of the current function make it hard for the players to use. By examining the groups in detail, new tools and functions could be developed, for example, to have a submenu for fragments on the design panel.

6 CONCLUSION

In this study, we compared comments from citizen and expert scientists using the think-aloud method. We found that citizen and expert scientists provide different information. The main reason for the difference is that expert scientists provide more background knowledge. Other categories do not show a significant discrepancy for different levels of domain knowledge experience.

A detailed analysis of the comments led to actionable recommendations about the interface and background information that should be supplied with the game for use by citizen scientists. In particular, drug design for proteins is challenging and citizen scientists, even

those with some background in chemistry, often expressed confusion about the processes that they were designing around. This finding suggests that future revisions of the game incorporate more background knowledge described by the expert scientists to reduce the confusion. Likewise, the user interface presented challenges to citizen and expert scientists alike. Incorporating the feedback from this study should improve the user experience overall.

This study is limited because it relies on the counts of comments to find distinguished issues and prioritize redesign efforts. However, counts of comments do not necessarily correspond to the severity of problems or the priority of refactoring. We are in the process of analyzing telemetry data (mouse movements, clicks, and keystrokes) from these user studies that may shed insight into the interface. When combined with our think-aloud analysis, we expect that further improvements in the FDDM interface can be made.

This work, however, provides a theoretical foundation of the difference between two main player groups in citizen science games during a usability study. An analysis of the difference also sheds lights on the refactoring workflow considering the difference for other researchers and designers in the fields

ACKNOWLEDGMENTS

The authors thank the reviewers for their constructive comments, Josh Miller for feedback and support, and participants for their time and effort. This material is based in part upon work supported by the National Science Foundation under Grant No. 1629811.

REFERENCES

Rick Bonney, Jennifer L. Shirk, Tina B. Phillips, Andrea Wiggins, Heidi L. Ballard, Abraham J. Miller-Rushing, and Julia K. Parrish. 2014. Next Steps for Citizen Science. *Science* 343, 6178 (2014), 1436–1437. https://doi.org/10.1126/science.1251554 arXiv:https://science.sciencemag.org/content/343/6178/1436.full.pdf

Anne Bowser, Derek Hansen, Yurong He, Carol Boston, Matthew Reid, Logan Gunnell, and Jennifer Preece. 2013. Using Gamification to Inspire New Citizen Science Volunteers. In *Proceedings of the First International Conference on Gameful Design, Research, and Applications* (Toronto, Ontario, Canada) *(Gamification '13)*. Association for Computing Machinery, New York, NY, USA, 18–25. https://doi.org/10.1145/2583008.2583011

Seth Cooper, Firas Khatib, Adrien Treuille, Janos Barbero, Jeehyung Lee, Michael Beenen, Andrew Leaver-Fay, David Baker, Zoran Popović, et al. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307 (2010), 756–760.

K. Crowston and N. R. Prestopnik. 2013. Motivation and Data Quality in a Citizen Science Game: A Design Science Evaluation. In *2013 46th Hawaii International Conference on System Sciences*. 450–459. https://doi.org/10.1109/HICSS.2013.413

J A DiMasi, L Feldman, A Seckler, and A Wilson. 2010. Trends in Risks Associated With New Drug Development: Success Rates for Investigational Drugs. *Clinical Pharmacology & Therapeutics* 87, 3 (2010), 272–277. https://doi.org/10.1038/clpt.2009.295 arXiv:https://ascpt.onlinelibrary.wiley.com/doi/pdf/10.1038/clpt.2009.295

Christopher B Eiben, Justin B Siegel, Jacob B Bale, Seth Cooper, Firas Khatib, Betty W Shen, Barry L Stoddard, Zoran Popovic, and David Baker. 2012. Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. *Nature biotechnology* 30, 2 (2012), 190–192.

K Anders Ericsson and Herbert A Simon. 1984. *Protocol analysis: Verbal reports as data*. the MIT Press.

Robert Heck, Oana Vuculescu, Jens Jakob Sørensen, Jonathan Zoller, Morten G. Andreasen, Mark G. Bason, Poul Ejlertsen, Ottó Elíasson, Pinja Haikka, Jens S. Laustsen, Lærke L. Nielsen, Andrew Mao, Romain Müller, Mario Napolitano, Mads K. Pedersen, Aske R. Thorsen, Carsten Bergenholtz, Tommaso Calarco, Simone Montangero, and Jacob F. Sherson. 2018. Remote optimization of an ultracold atoms experiment by experts and citizen scientists. *Proceedings of the National Academy of Sciences* 115, 48 (2018), E11231–E11237. https://doi.org/10.1073/pnas.1716869115 arXiv:https://www.pnas.org/content/115/48/E11231.full.pdf

Scott Horowitz, Brian Koepnick, Raoul Martin, Agnes Tymieniecki, Amanda A Winburn, Seth Cooper, Jeff Flatten, David S Rogawski, Nicole M Koropatkin, Tsinatkeab T Hailu, et al. 2016. Determining crystal structures through crowdsourcing and coursework. *Nature communications* 7, 1 (2016), 1–11.

Ioanna Iacovides, Charlene Jennett, Cassandra Cornish-Trestrail, and Anna L. Cox. 2013. Do Games Attract or Sustain Engagement in Citizen Science? A Study of Volunteer Motivations. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13)*. Association for Computing Machinery, New York, NY, USA, 1101–1106. https://doi.org/10.1145/2468356.2468553

Monique W.M. Jaspers. 2009. A comparison of usability methods for testing interactive health technologies: Methodological aspects and empirical evidence. *International Journal of Medical Informatics* 78, 5 (2009), 340 – 353. https://doi.org/10.1016/j.ijmedinf.2008.10.002

Monique W.M. Jaspers, Thiemo Steen, Cor van den Bos, and Maud Geenen. 2004. The think aloud method: a guide to user interface design. *International Journal of Medical Informatics* 73, 11 (2004), 781 – 795. https://doi.org/10.1016/j.ijmedinf.2004.08.003

Victor Jupp. 2006. *The Sage dictionary of social research methods*. Sage.

Firas Khatib, Seth Cooper, Michael D. Tyka, Kefan Xu, Ilya Makedon, Zoran Popović, David Baker, and Foldit Players. 2011a. Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences* 108, 47 (2011), 18949–18953. https://doi.org/10.1073/pnas.1115898108

Firas Khatib, Frank DiMaio, Seth Cooper, Maciej Kazmierczyk, Miroslaw Gilski, Szymon Krzywda, Helena Zabranska, Iva Pichova, James Thompson, Zoran Popović, et al. 2011b. Crystal structure of a monomeric retroviral protease solved by protein folding game players. *Nature structural & molecular biology* 18, 10 (2011), 1175–1177.

Brian Koepnick, Jeff Flatten, Tamir Husain, Alex Ford, Daniel-Adriano Silva, Matthew J Bick, Aaron Bauer, Gaohua Liu, Yojiro Ishida, Alexander Boykov, et al. 2019. De novo protein design by citizen scientists. *Nature* 570, 7761 (2019), 390–394.

E. Korpela, D. Werthimer, D. Anderson, J. Cobb, and M. Leboisky. 2001. SETI@home-massively distributed computing for SETI. *Computing in Science Engineering* 3, 1 (Jan 2001), 78–83. https://doi.org/10.1109/5992.895191

Andrew Leaver-Fay, Michael Tyka, Steven M. Lewis, Oliver F. Lange, James Thompson, Ron Jacak, Kristian W. Kaufman, P. Douglas Renfrew, Colin A. Smith, Will Sheffler, Ian W. Davis, Seth Cooper, Adrien Treuille, Daniel J. Mandell, Florian Richter, Yih-En Andrew Ban, Sarel J. Fleishman, Jacob E. Corn, David E. Kim, Sergey Lyskov, Monica Berrondo, Stuart Mentzer, Zoran Popović, James J. Havranek, John Karanicolas, Rhiju Das, Jens Meiler, Tanja Kortemme, Jeffrey J. Gray, Brian Kuhlman, David Baker, and Philip Bradley. 2011. Chapter nineteen - Rosetta3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules. In *Computer Methods, Part C*, Michael L. Johnson and Ludwig Brand (Eds.). Methods in Enzymology, Vol. 487. Academic Press, 545 – 574. https://doi.org/10.1016/B978-0-12-381270-4.00019-6

Jens Meiler and David Baker. 2006. ROSETTALIGAND: Protein–small molecule docking with full side-chain flexibility. *Proteins: Structure, Function, and Bioinformatics* 65, 3 (2006), 538–548. https://doi.org/10.1002/prot.21086

Josh Aaron Miller, Vivian Lee, Seth Cooper, and Magy Seif El-Nasr. 2019. Large-Scale Analysis of Visualization Options in a Citizen Science Game. In *Extended Abstracts of the Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts* (Barcelona, Spain) *(CHI PLAY '19 Extended Abstracts)*. Association for Computing Machinery, New York, NY, USA, 535–542. https://doi.org/10.1145/3341215.3356274

Garrett M. Morris, Ruth Huey, William Lindstrom, Michel F. Sanner, Richard K. Belew, David S. Goodsell, and Arthur J. Olson. 2009. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry* 30, 16 (2009), 2785–2791. https://doi.org/10.1002/jcc.21256

Michael Shirts and Vijay S. Pande. 2000. Screen Savers of the World Unite! *Science* 290, 5498 (2000), 1903–1904. https://doi.org/10.1126/science.290.5498.1903

Jonathan Silvertown. 2009. A new dawn for citizen science. *Trends in Ecology & Evolution* 24, 9 (2009), 467 – 471. https://doi.org/10.1016/j.tree.2009.03.017

Kristin Siu, Alexander Zook, and Mark O. Riedl. 2017. A Framework for Exploring and Evaluating Mechanics in Human Computation Games. In *Proceedings of the 12th International Conference on the Foundations of Digital Games* (Hyannis, Massachusetts) *(FDG '17)*. Association for Computing Machinery, New Yorkk NYk USA, Article 38, 4 pages. https://doi.org/10.1145/3102071.3106344

Joanna Stanczyk, Caroline Ospelt, and Steffen Gay. 2008. Is there a future for small molecule drugs in the treatment of rheumatic diseases? *Current opinion in rheumatology* 20, 3 (2008), 257–262.

Brian L. Sullivan, Christopher L. Wood, Marshall J. Iliff, Rick E. Bonney, Daniel Fink, and Steve Kelling. 2009. eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* 142, 10 (2009), 2282 – 2292. https://doi.org/10.1016/j.biocon.2009.05.006

Marcel L. Verdonk, Jason C. Cole, Michael J. Hartshorn, Christopher W. Murray, and Richard D. Taylor. 2003. Improved protein–ligand docking using GOLD. *Proteins: Structure, Function, and Bioinformatics* 52, 4 (2003), 609–623. https://doi.org/10.1002/prot.10465 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.10465

L. von Ahn. 2006. Games with a purpose. *Computer* 39, 6 (June 2006), 92–94. https://doi.org/10.1109/MC.2006.196

Luis Von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Commun. ACM* 51, 8 (2008), 58–67.

SUPPLEMENTARY MATERIAL TO FOLDIT DRUG DESIGN GAME USABILITY STUDY: COMPARISON OF CITIZEN AND EXPERT SCIENTISTS

A Relaxed Think-aloud Scheme

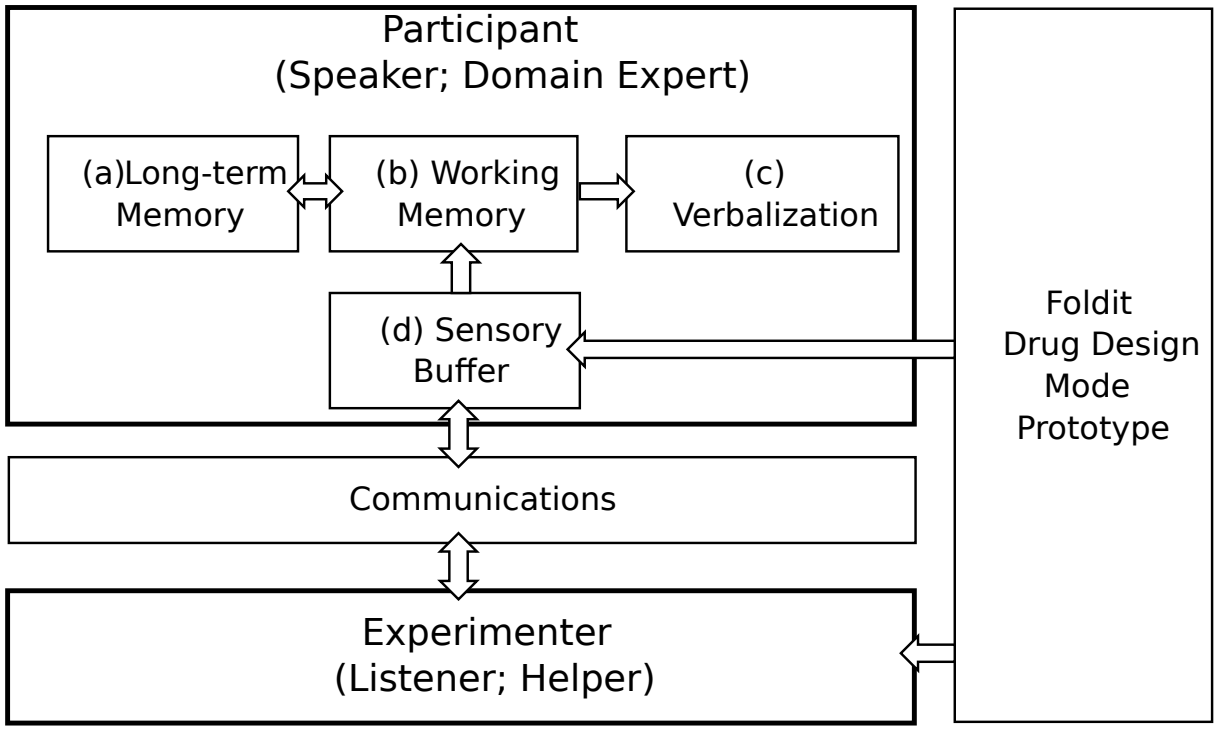

Figure S1: The participant (c) verbalizes thoughts in the working memory (b) using knowledge from long-term memory (a) and information from the sensory buffer (d). The experimenter remains a main listener or helper role whereas the participant is in the role of the main speaker and domain expert. The experimenter communicates with the participant without stopping the gameplay. The experimenter also observes the prototype.

Relation between Unique Descriptors and Categories

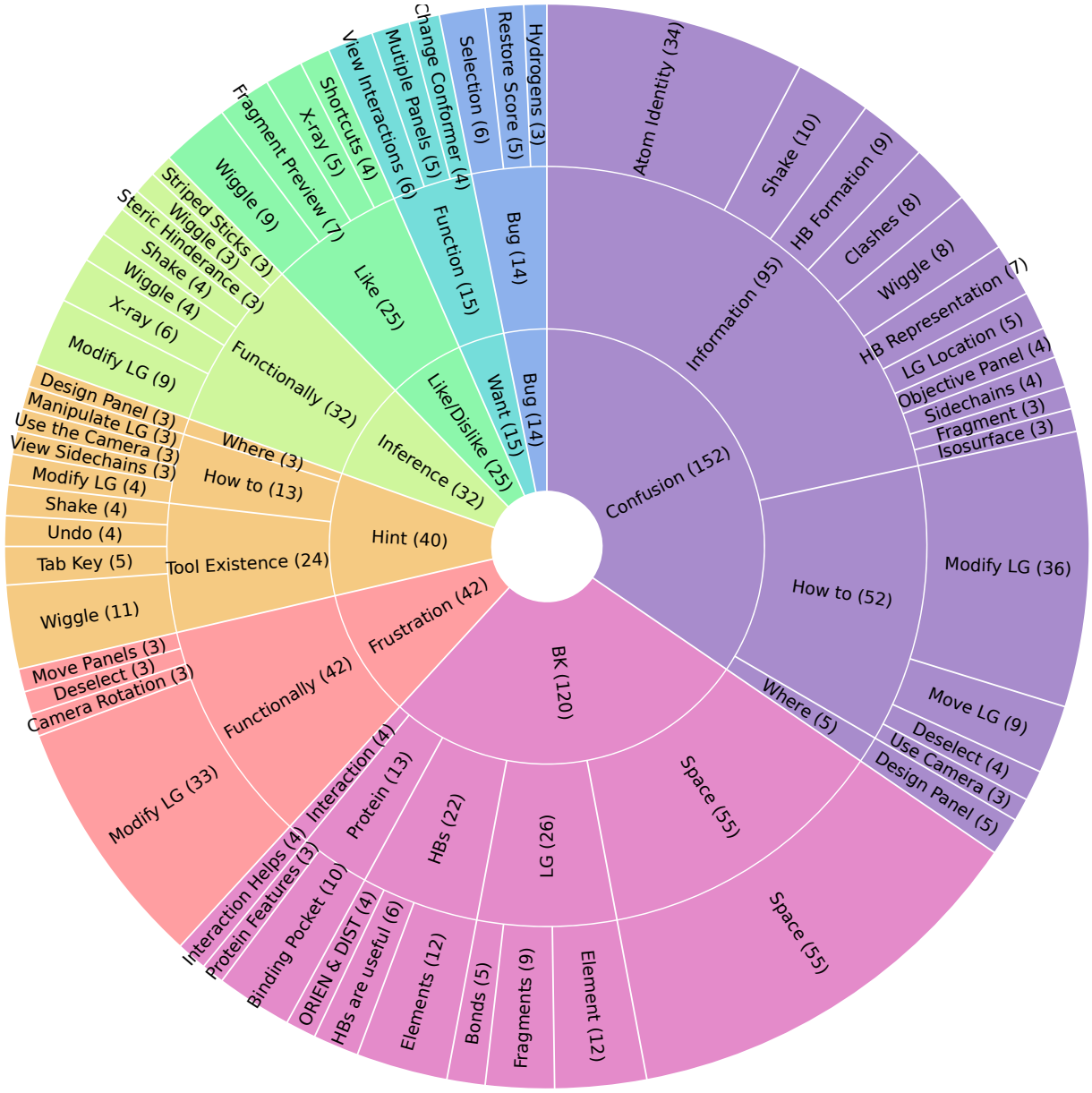

Figure S2: Sunburst diagram of Unique Descriptions (UDs) for similar comments. UDs are grouped into three-level hierachy, with the center being the highest level and the outer being the leaf level. Numbers in parenthesis denote the number of comments in that UD and areas are proportional to counts. Comments from less than three participants were excluded from this diagram and further analysis. 'HB' stands for hydrogen bonds. 'LG' stands for ligand. 'ORIEN & DIST' stands for orientation and distance. 'BK' stands for background knowledge (Diagram based on https://observablehq.com/@d3/sunburst?collection=@d3/d3-hierarchy).

Example User Comments

Table S1: Example user comments for each category.

Category	Example Comments
confusion	• "I don't know how to select a molecule and put it where I want it to go." • "I guess what I am wondering then is when you are in the mode that wiggles, if new hydrogen bonds are formed, will it show those in that mode? " • "Oh, I am not sure what I did."
background knowledge	• "Hmm, let's try something that can be a hydrogen bond donor or hydrogen bond acceptor." • "Hydrogen bond is certain distance away. " • "It looks like I have a decent amount of space in the pocket."
want	• "May I have a little comment … when things are disabled, there is no things, like notice." • "I wonder if there is a way to see the distance between two atoms." • "it would be good to put an option like 'fill out the valence shell'."
like/dislike	• "That's a cool feature." • "It's definitely harder to see things when you don't have the X-ray on."(X-ray is a tool in the view panel). • "I don't like that, when I click inside, you have to click over here to move it."
hint	• "There is also an undo panel if you want to undo this." • "Most of time, after modifying it, you may want to use wiggle." • "When you hover your mouse over the sidechain and press tab, it shows the info"
bug	• "There seems to be something wrong." • "What happened here." • "I don't know what's going on with that one."
inference	• "Oh! I guess you select the molecule, and then you click on one of these." • "I am not sure what this circle means. (playing with the X-ray circle) ...oh I know now; this means more out ...um ...in front of the protein structure." • "Ok, so that puts it in the most ideal conformation, is that what it does?"
frustration	• "(sigh) ok whatever, I am gonna see if I can move onto something else." • "I'm frustrated because I cannot figure out, like I have bonds over here, but I cannot figure out how to get interactions in other places." • "I'm trying to attach an atom here, but I can't."

Raw Data

Table S2: Data Points for All Categories

Category	Play Group	# Comments from Each Player											
confusion	citizen	8	11	17	6	14	20	23	15	8	15	2	12
	expert	19	2	8	14	3	13	11	5	1	17		
background knowledge	citizen	6	2	8	1	3	5	5	-	1	1	6	0
	expert	11	27	23	19	11	7	1	13	6	5		
want	citizen	-	2	8	4	1	-	2	1	1	-	1	6
	expert	4	-	-	1	1	14	2	3	-	4		
like/dislike	citizen	2	-	2	5	3	6	3	1	5	1	1	2
	expert	1	1	-	6	-	17	13	5	1	2		
hint	citizen	2	1	2	3	4	6	6	11	7	3	2	3
	expert	1	-	2	10	4	5	-	2	1	5		
bug	citizen	1	1	1	1	3	1	2	1	1	-	-	1
	expert	-	1	2	1	-	2	-	2	-	-		
inference	citizen	7	5	1	7	4	3	6	9	12	1	1	3
	expert	6	2	4	8	1	2	1	-	3	7		
frustration	citizen	3	3	1	1	1	13	1	3	1	3	-	1
	expert	8	2	5	5	-	1	1	2	1	-		

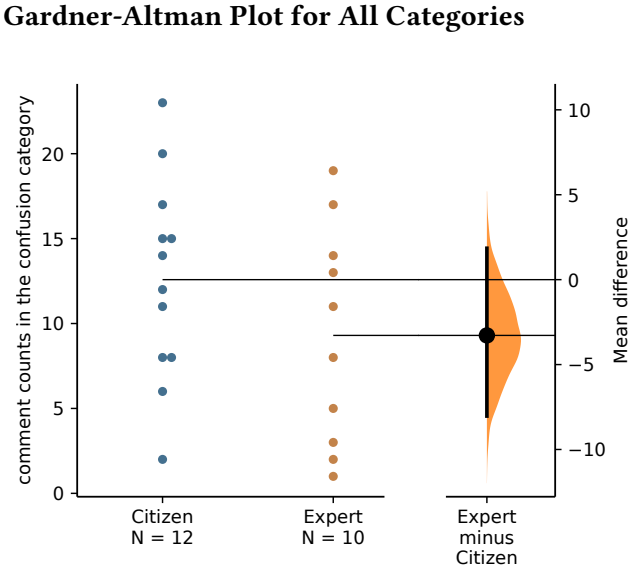

Figure S3: t-test between citizen and expert scientists in the confusion category.

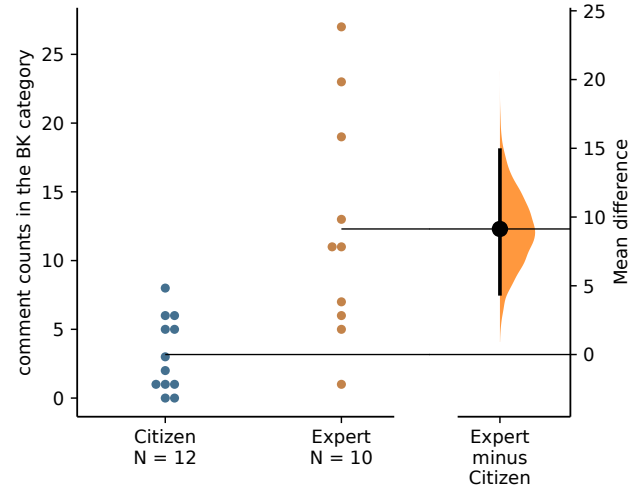

Figure S4: t-test between citizen and expert scientists in the background knowledge category.

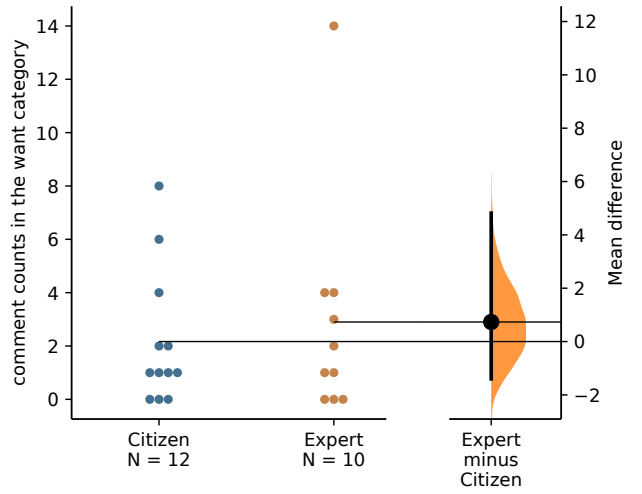

Figure S5: t-test between citizen and expert scientists in the want category.

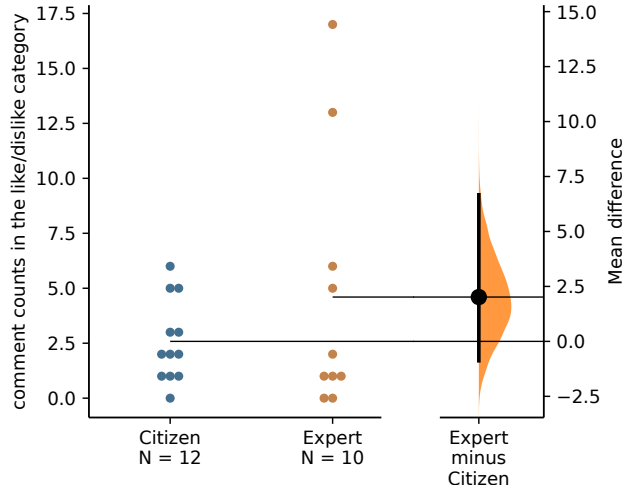

Figure S6: t-test between citizen and expert scientists in the like/dislike category.

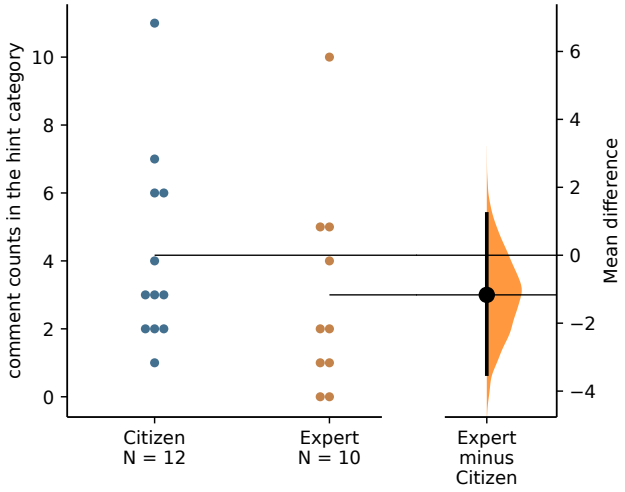

Figure S7: t-test between citizen and expert scientists in the hint category.

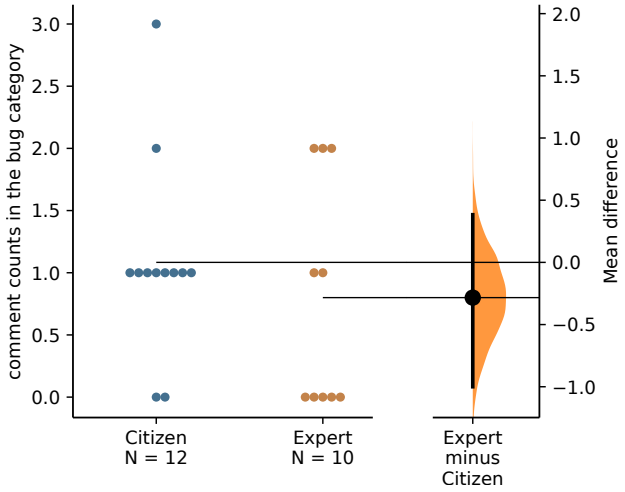

Figure S8: t-test between citizen and expert scientists in the bug category.

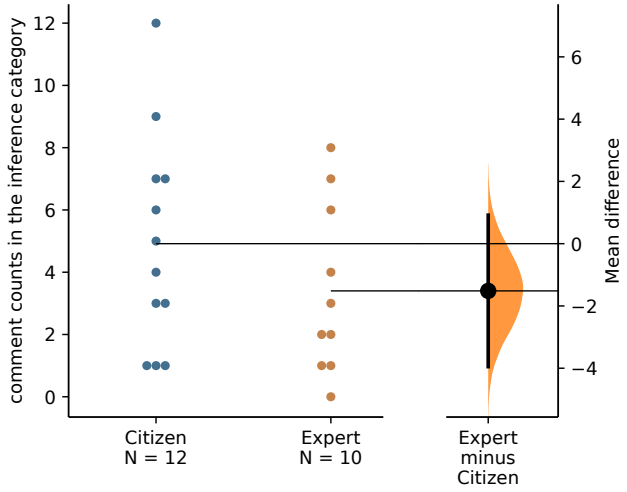

Figure S9: t-test between citizen and expert scientists in the inference category.

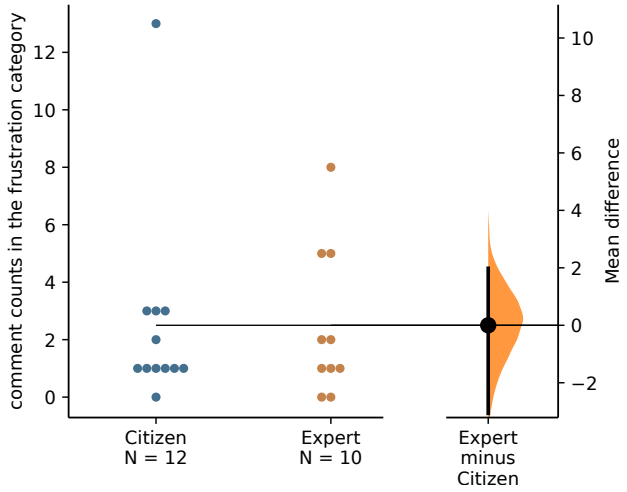

Figure S10: t-test between citizen and expert scientists in the frustration category.

Results for RQ1

Table S3: Total number of comments (citizen, expert), p-values, mean differences and 95% confidence intervals (95% CI) for all categories. The number of comments is summed over all participants in the group.

category	# comments	p-value	mean diff.	95% CI
confusion	(151, 93)	0.226	-3.28	[-8.03,1.85]
background knowledge	(38, 123)	< 0.001	9.13	[4.42,14.9]
want	(26, 29)	0.632	0.733	[-1.4,4.82]
like/dislike	(31, 46)	0.315	2.02	[-0.883,6.67]
hint	(50, 30)	0.359	-1.17	[-3.5,1.22]
bug	(13, 8)	0.316	-0.283	[-1.0, 0.383]
inference	(59, 34)	0.257	-1.52	[-3.95,0.917]
frustration	(30, 25)	0.951	0.0	[-3.07,1.98]

Comment Distribution

Table S4: Comment distribution for main subgroups in the 'chemistry concept' group (like/dislike, hint, bug, frustration columns are omitted due to zero comment counts).

Subgroup	Component	Player Group	confusion	background knowledge	want	inference
general chemistry	hydrogen bond	citizen	7	4	-	3
		expert	2	18	-	-
	atoms	citizen	23	4	-	-
		expert	11	8	-	-
	fragment	citizen	-	-	-	-
		expert	-	9	-	-
	bond	citizen	-	2	-	-
		expert	-	3	-	-
drug design	ligand	citizen	3	-	-	-
		expert	2	-	4	-
	space	citizen	-	20	-	-
		expert	-	35	-	-
	interaction	citizen	-	2	-	-
		expert	-	2	-	-

Table S5: Comment distribution for main subgroups in the 'game element' group (background knowledge, want columns are omitted due to zero comments counts).

Subgroup	Component	Player Group	confusion	like/dislike	hint	bug	inference	frustration
modes panel	ligand design	citizen	33	5/-	3	-	3	19
		expert	15	2/-	4	-	6	17
	pull mode	citizen	3	-	-	-	-	-
		expert	6	-	3	-	-	-
actions panel	shake	citizen	5	-	4	-	4	-
		expert	5	-	-	-	-	-
	wiggle	citizen	5	4/-	10	5	5	-
		expert	3	5/-	1	1	2	-
view panel	show clashes	citizen	8	-	-	-	5	-
		expert	-	-	-	-	1	-
	show bonds	citizen	4	-	-	-	-	-
		expert	3	-	-	-	-	-
	show isosurface	citizen	2	-	-	-	-	-
		expert	1	-	-	-	-	-
	X-ray tunnel	citizen	-	2/-	-	-	1	-
		expert	-	3/-	-	-	2	-
	view hydrogens	citizen	-	-	-	2	-	-
		expert	-	-	-	1	-	-
	view sidechains	citizen	-	-	-	-	-	-
		expert	-	-	3	-	-	-

Screenshots of the Panels

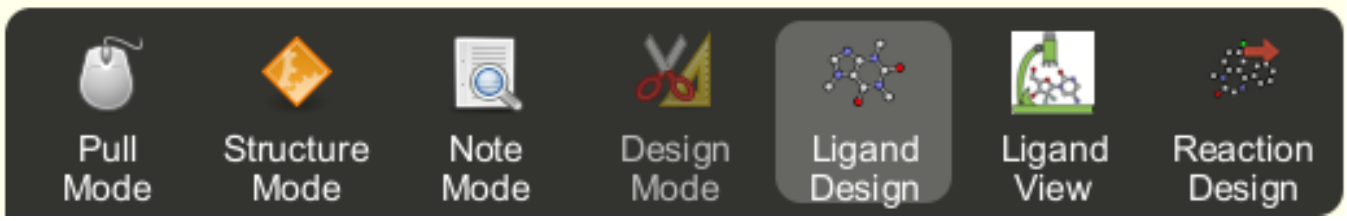

Figure S11: Modes panel. The pull mode allows the player to manually manipulate the ligand, i.e. rotate the ligand, rotate the bond, translation. The ligand design opens small molecule design palette shown in Figure S14. The ligand view shows properties such as weight, cLogP, polar surface area, as well as shows visual presentations such as cation-pi, pi-pi interactions and isosurface around the ligand. The reaction design is under development at the time of writing. Other icons are not related to this study.

Figure S12: Actions Panel. Shake and wiggle are two most important tools. For the protein, Wiggle Backbone only applies to backbones while Wiggle Sidechains applies only to sidechains. Both moves the ligand. Wiggle All is the sum of both Wiggle Backbone and Wiggle Sidechains. Other tools are not relevant in this study.

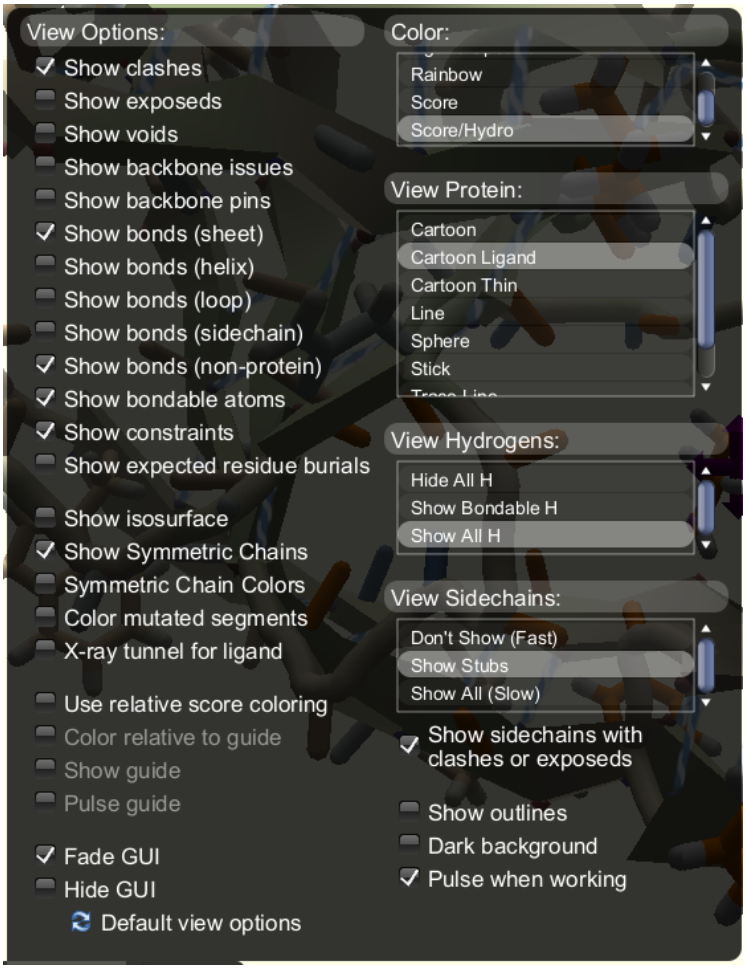

Figure S13: View panel. This is study uses an advanced interface mode. There are much fewer options if not in an advanced interface mode.

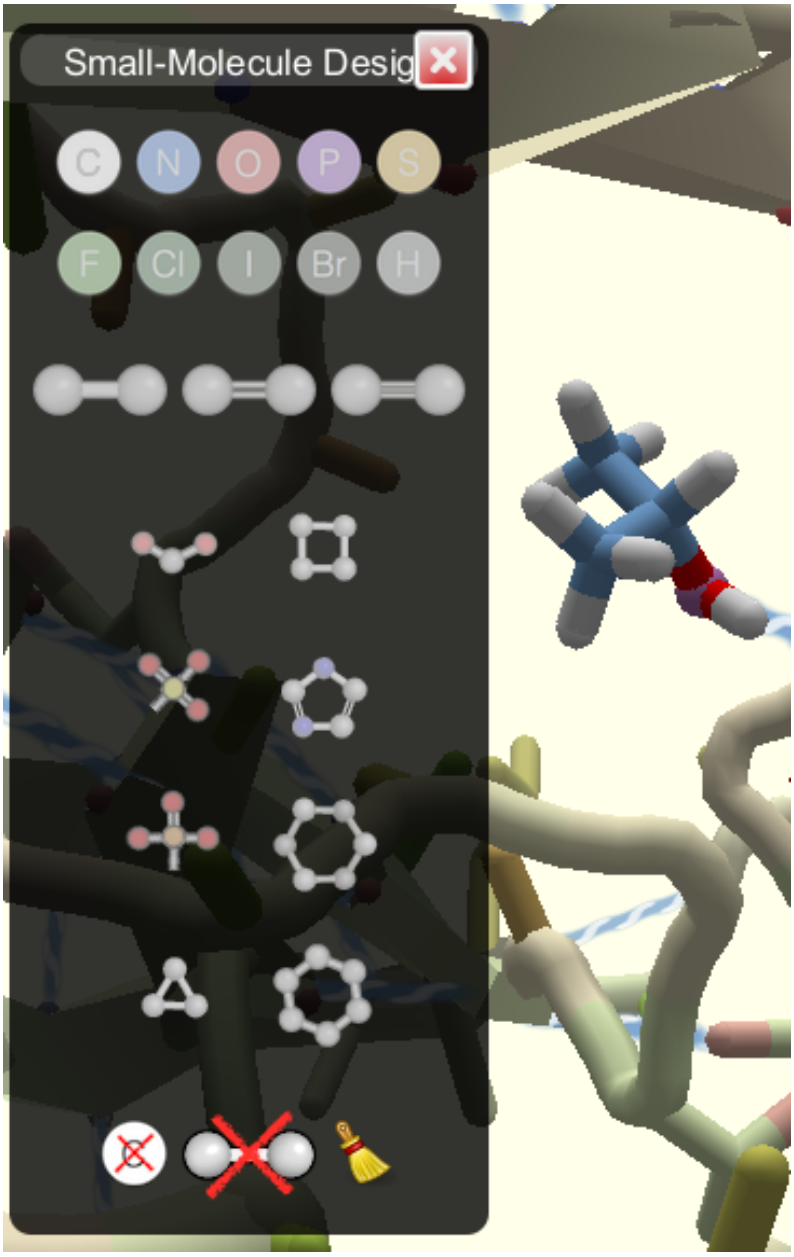

Figure S14: Small molecule design palette that is accessed from the ligand design on modes panel. The top region is for single element. Next come three types of bonds: single bond, double bond, triple bond. It follows fragment section. The icons at the last row are delete atom, delete bond and clear selection. The orange-white structure to the right of the design palette is the ligand to be modified.

Current Way of Modifying the Ligand

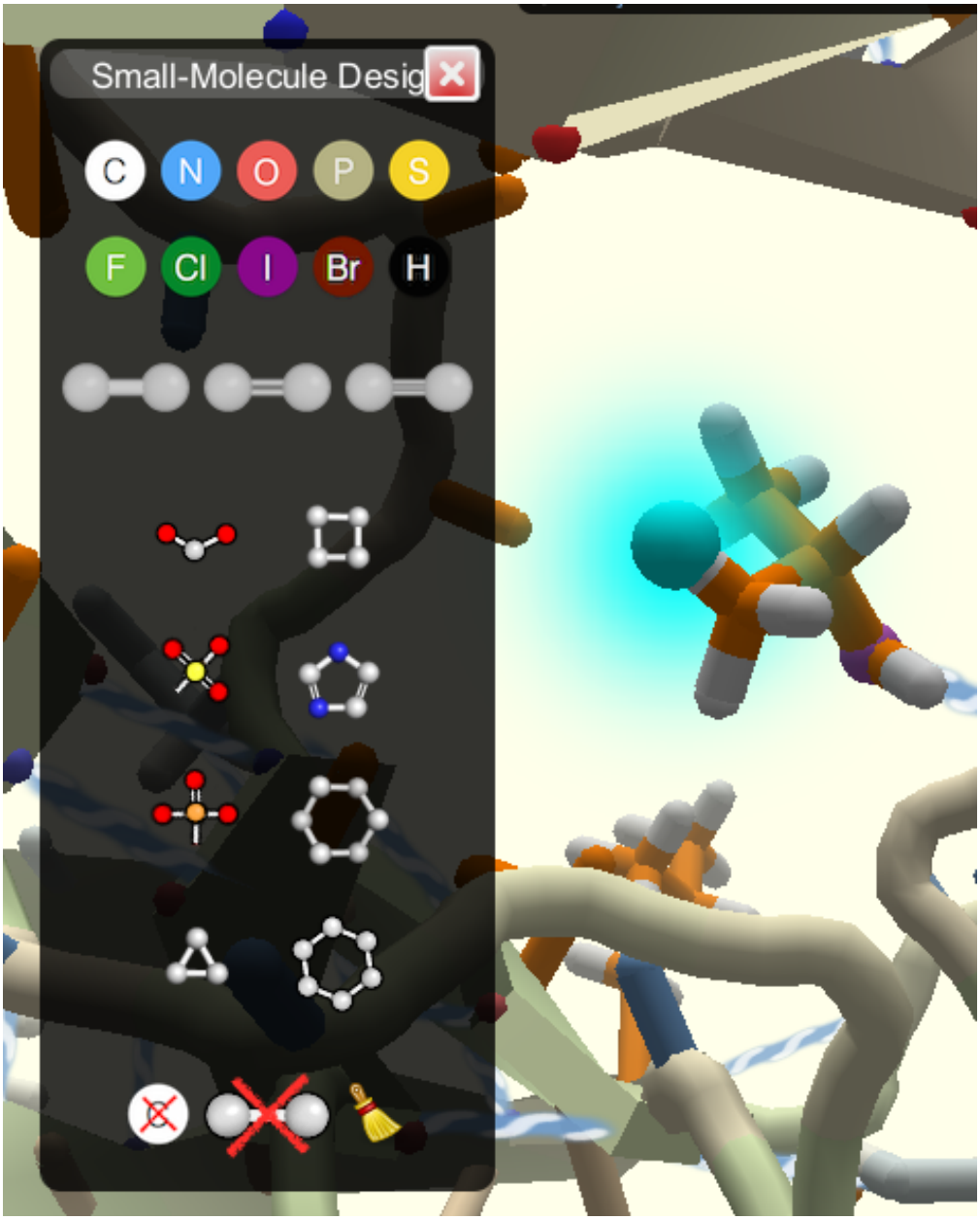

Figure S15: Selection of atom by left clicking.

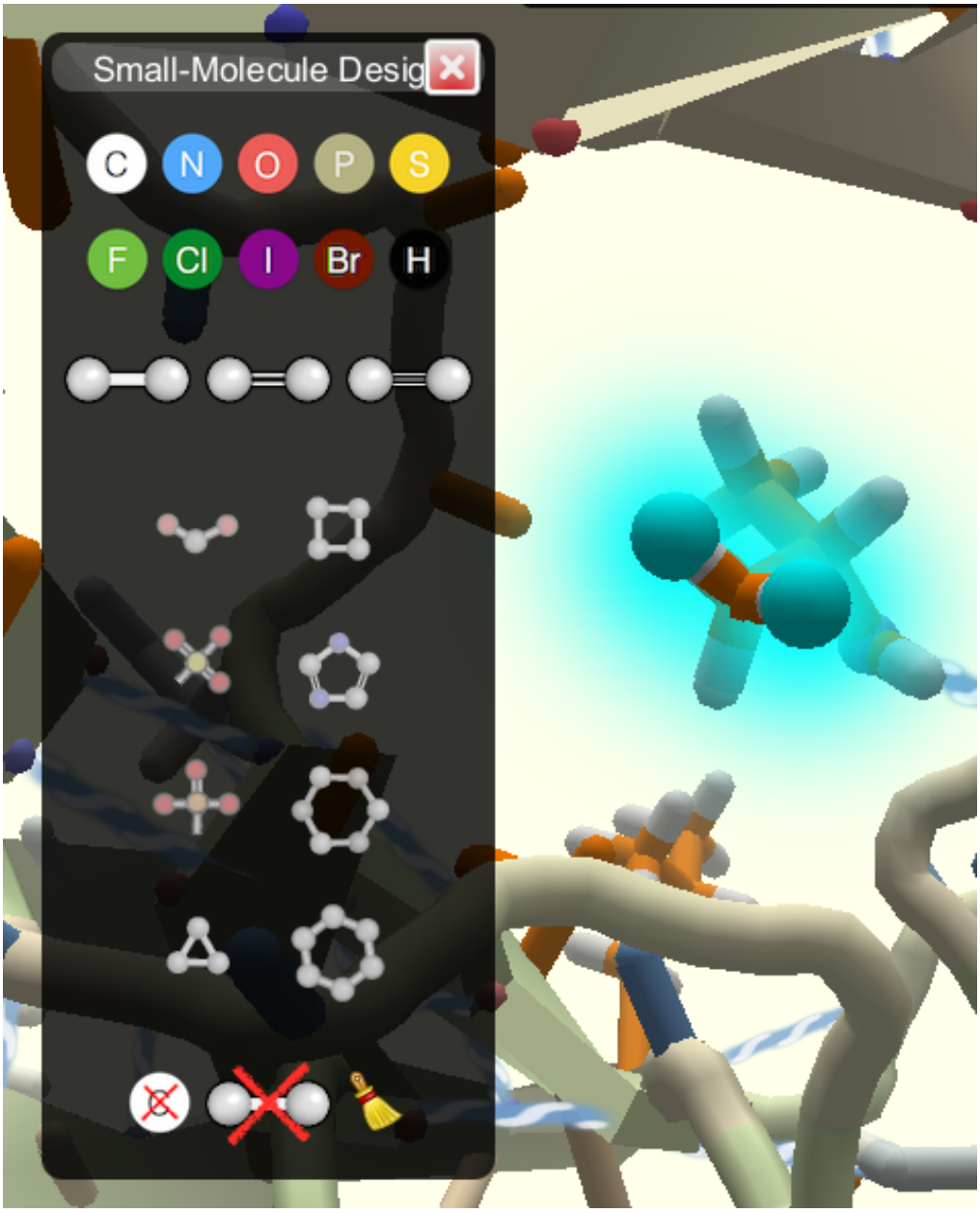

Figure S16: Selecting more than one atom activates bond addition on design panel.

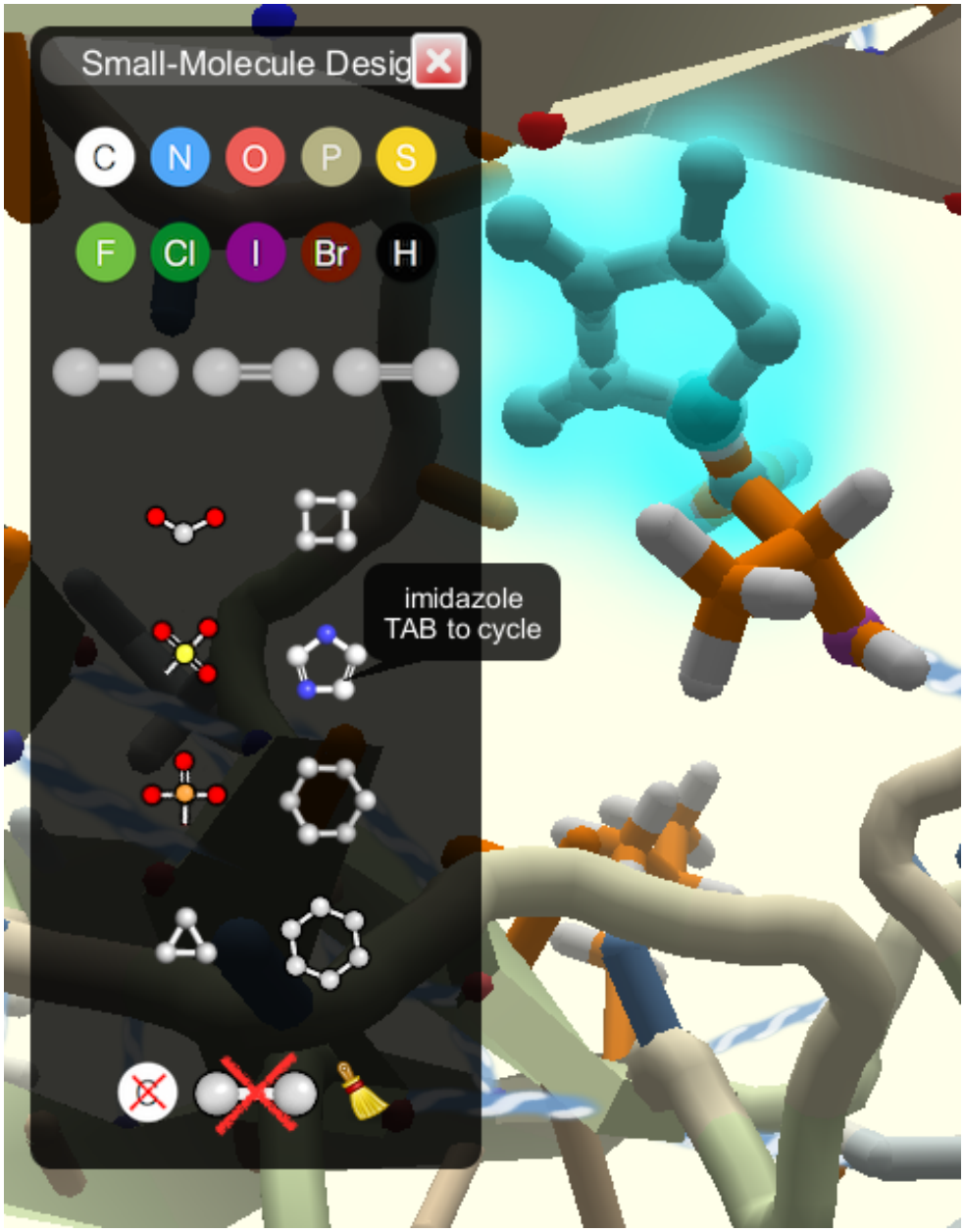

Figure S17: Hovering mouse over fragment will show a preview on the ligand.