[Skip to content](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare/e2325c81...master" \l "content-body)

Toggle navigation

Top of Form

This project

Bottom of Form

* [C32c205abedc5e6cc2214e5c6c82eb68?s=52&d=identicon](https://spotlite.nih.gov/debug22)

[ncats](https://spotlite.nih.gov/ncats) / [ScaffoldAnalytics](https://spotlite.nih.gov/ncats/scaffoldanalytics)

[](https://spotlite.nih.gov/)

Toggle navigation

Toggle navigation pinning

* [Projects](https://spotlite.nih.gov/dashboard/projects)
* [Activity](https://spotlite.nih.gov/dashboard/activity)
* [Groups](https://spotlite.nih.gov/dashboard/groups)
* [Milestones](https://spotlite.nih.gov/dashboard/milestones)
* [Issues0](https://spotlite.nih.gov/dashboard/issues?assignee_id=14)
* [Merge Requests0](https://spotlite.nih.gov/dashboard/merge_requests?assignee_id=14)
* [Snippets](https://spotlite.nih.gov/dashboard/snippets)

[About GitLab CE](https://spotlite.nih.gov/help)

* [Project](https://spotlite.nih.gov/ncats/scaffoldanalytics)
* [Activity](https://spotlite.nih.gov/ncats/scaffoldanalytics/activity)
* [**Repository**](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/master)
* [Pipelines](https://spotlite.nih.gov/ncats/scaffoldanalytics/pipelines)
* [Graphs](https://spotlite.nih.gov/ncats/scaffoldanalytics/graphs/master)
* [Issues 0](https://spotlite.nih.gov/ncats/scaffoldanalytics/issues)
* [Merge Requests 0](https://spotlite.nih.gov/ncats/scaffoldanalytics/merge_requests)
* [Wiki](https://spotlite.nih.gov/ncats/scaffoldanalytics/wikis/home)
* [Files](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/master)
* [Commits](https://spotlite.nih.gov/ncats/scaffoldanalytics/commits/master)
* [Network](https://spotlite.nih.gov/ncats/scaffoldanalytics/network/master)
* [**Compare**](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare?from=master&to=master)
* [Branches](https://spotlite.nih.gov/ncats/scaffoldanalytics/branches)
* [Tags](https://spotlite.nih.gov/ncats/scaffoldanalytics/tags)

Top of Form

from

e2325c81

...

to

master

   Compare

Bottom of Form

Commits (18)

* [](mailto:rajarshi_guha@vrtx.com)

[**reworked discussion section**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/b8d9f854cdffd9429d57ccdeb5876e59665e58d0)

[Rajarshi Guha](mailto:rajarshi_guha@vrtx.com) committed 2 weeks ago

[**b8d9f854**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/b8d9f854cdffd9429d57ccdeb5876e59665e58d0) [Browse Files](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/b8d9f854cdffd9429d57ccdeb5876e59665e58d0)

* [](mailto:rajarshi_guha@vrtx.com)

[**tightened up discussion**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/4e6b2adb1e003034fae398fd7e17da7e68ca5833)

[Rajarshi Guha](mailto:rajarshi_guha@vrtx.com) committed 2 weeks ago

[**4e6b2adb**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/4e6b2adb1e003034fae398fd7e17da7e68ca5833) [Browse Files](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/4e6b2adb1e003034fae398fd7e17da7e68ca5833)

* [](mailto:rajarshi_guha@vrtx.com)

[**reordered NCATS affiliations. Added current address for myself**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/fbfe0178e2154fd94ad43e20ee5da1da4d7c9189)

[Rajarshi Guha](mailto:rajarshi_guha@vrtx.com) committed 2 weeks ago

[**fbfe0178**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/fbfe0178e2154fd94ad43e20ee5da1da4d7c9189) [Browse Files](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/fbfe0178e2154fd94ad43e20ee5da1da4d7c9189)

* [](mailto:rajarshi_guha@vrtx.com)

[**some minor edits**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/c4f1286a971c26d01216921313dcae90d5ca3875)

[Rajarshi Guha](mailto:rajarshi_guha@vrtx.com) committed 2 weeks ago

[**c4f1286a**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/c4f1286a971c26d01216921313dcae90d5ca3875) [Browse Files](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/c4f1286a971c26d01216921313dcae90d5ca3875)

* [](mailto:dbandyo3@its.jnj.com)

[**modified details figure to remove SNG**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/82f46ab4dabbcca378fa7355912f0ea8c2da7f6b)

[NA\DBandyo3](mailto:dbandyo3@its.jnj.com) committed 2 weeks ago

[**82f46ab4**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/82f46ab4dabbcca378fa7355912f0ea8c2da7f6b) [Browse Files](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/82f46ab4dabbcca378fa7355912f0ea8c2da7f6b)

* [](mailto:dbandyo3@its.jnj.com)

[**Progress on paper as of Sat 7/28 morning. Stopped around Visualization section. …**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/b477c411dd434dd6d3b6e1527d4ca71c8dc87e3e) ...

[NA\DBandyo3](mailto:dbandyo3@its.jnj.com) committed 2 weeks ago

[**b477c411**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/b477c411dd434dd6d3b6e1527d4ca71c8dc87e3e) [Browse Files](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/b477c411dd434dd6d3b6e1527d4ca71c8dc87e3e)

* [](mailto:dbandyo3@its.jnj.com)

[**removed 'Size by:' from corner of spotviz\_scafpie\_tooltip.png**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/ef9bfef05e59aff3d3ab7df831219fd852d813b0)

[NA\DBandyo3](mailto:dbandyo3@its.jnj.com) committed 2 weeks ago

[**ef9bfef0**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/ef9bfef05e59aff3d3ab7df831219fd852d813b0) [Browse Files](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/ef9bfef05e59aff3d3ab7df831219fd852d813b0)

* [](mailto:dbandyo3@its.jnj.com)

[**Progress on paper as of Sat 7/28 1pm. Reached till Results section.**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/f586e666e8e2fce2620c4bcaf8ed6858ce1b561c)

[NA\DBandyo3](mailto:dbandyo3@its.jnj.com) committed 2 weeks ago

[**f586e666**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/f586e666e8e2fce2620c4bcaf8ed6858ce1b561c) [Browse Files](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/f586e666e8e2fce2620c4bcaf8ed6858ce1b561c)

* [](mailto:dbandyo3@its.jnj.com)

[**Few more minor edits**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/38319a1a0ef9a73ef8b3c46a5c3b7010a6e16e72)

[NA\DBandyo3](mailto:dbandyo3@its.jnj.com) committed 2 weeks ago

[**38319a1a**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/38319a1a0ef9a73ef8b3c46a5c3b7010a6e16e72) [Browse Files](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/38319a1a0ef9a73ef8b3c46a5c3b7010a6e16e72)

* [](mailto:dbandyo3@its.jnj.com)

[**Reached till quantitative comparison... fixed text and one figure**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/52b306d67af00776750451b820dd47de4917edae)

[NA\DBandyo3](mailto:dbandyo3@its.jnj.com) committed 2 weeks ago

[**52b306d6**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/52b306d67af00776750451b820dd47de4917edae) [Browse Files](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/52b306d67af00776750451b820dd47de4917edae)

* [](mailto:dbandyo3@its.jnj.com)

[**First pass is complete - reached end of main paper including Discussion. Replace…**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/8b3336946201a056bc7c4a7fc8bc7c474f5f294d) ...

[NA\DBandyo3](mailto:dbandyo3@its.jnj.com) committed 2 weeks ago

[**8b333694**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/8b3336946201a056bc7c4a7fc8bc7c474f5f294d) [Browse Files](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/8b3336946201a056bc7c4a7fc8bc7c474f5f294d)

* [](mailto:dbandyo3@its.jnj.com)

[**Made some changes to SI as well as main text**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/f0c45decf80075b721aa79396c7746f547cfc0b1)

[NA\DBandyo3](mailto:dbandyo3@its.jnj.com) committed 2 weeks ago

[**f0c45dec**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/f0c45decf80075b721aa79396c7746f547cfc0b1) [Browse Files](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/f0c45decf80075b721aa79396c7746f547cfc0b1)

* [](mailto:dbandyo3@its.jnj.com)

[**Slide pack where statistics figure was updated to use transparent rectangles**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/0af9a8f398d43590f7018350299eaf882c16fca2)

[NA\DBandyo3](mailto:dbandyo3@its.jnj.com) committed a week ago

[**0af9a8f3**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/0af9a8f398d43590f7018350299eaf882c16fca2) [Browse Files](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/0af9a8f398d43590f7018350299eaf882c16fca2)

* [](mailto:dbandyo3@its.jnj.com)

[**Spotfire file where statistics plot was updated to use transparent rectangles**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/98e5e0adab6d36d55d3a8f87e0c0fd597d67b223)

[NA\DBandyo3](mailto:dbandyo3@its.jnj.com) committed a week ago

[**98e5e0ad**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/98e5e0adab6d36d55d3a8f87e0c0fd597d67b223) [Browse Files](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/98e5e0adab6d36d55d3a8f87e0c0fd597d67b223)

* [](mailto:rajarshi_guha@vrtx.com)

[**Updated abstract**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/7279a5e4463553ddb31ef1714ba029cdc33f88e4)

[Rajarshi Guha](mailto:rajarshi_guha@vrtx.com) committed about 3 hours ago

[**7279a5e4**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/7279a5e4463553ddb31ef1714ba029cdc33f88e4) [Browse Files](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/7279a5e4463553ddb31ef1714ba029cdc33f88e4)

* [](mailto:rajarshi_guha@vrtx.com)

[**More abstract edits. tightening text**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/3ab775dca3c6e6b57cc347ec8b844b74835fe831)

[Rajarshi Guha](mailto:rajarshi_guha@vrtx.com) committed about 2 hours ago

[**3ab775dc**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/3ab775dca3c6e6b57cc347ec8b844b74835fe831) [Browse Files](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/3ab775dca3c6e6b57cc347ec8b844b74835fe831)

* [](mailto:rajarshi_guha@vrtx.com)

[**reworking introduction**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/454999bb2f25ee7268b697c333d34f344fde5285)

[Rajarshi Guha](mailto:rajarshi_guha@vrtx.com) committed about 2 hours ago

[**454999bb**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/454999bb2f25ee7268b697c333d34f344fde5285) [Browse Files](https://spotlite.nih.gov/ncats/scaffoldanalytics/tree/454999bb2f25ee7268b697c333d34f344fde5285)

* [](mailto:rajarshi_guha@vrtx.com)

[**final edits**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/7fca65b4814833372e08489c48859581ecb70925)

[Rajarshi Guha](mailto:rajarshi_guha@vrtx.com) committed about 2 hours ago

[**7fca65b4**](https://spotlite.nih.gov/ncats/scaffoldanalytics/commit/7fca65b4814833372e08489c48859581ecb70925)

[Expand all](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare/e2325c81...master?expand_all_diffs=1) [Hide whitespace changes](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare/e2325c81...master?w=1)

[Inline](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare/e2325c81...master?view=inline)[Side-by-side](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare/e2325c81...master?view=parallel)

Showing [**10 changed files**](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare/e2325c81...master)with **0 additions** and **0 deletions**

[**SI/SI.pdf**](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare/e2325c81...master#eb471e773a04b3f193fbad58c600aed13ee0c362) 0 → 100644

[View file @7fca65b](https://spotlite.nih.gov/ncats/scaffoldanalytics/blob/7fca65b4814833372e08489c48859581ecb70925/SI/SI.pdf)

No preview for this file type

[**SI/SI.tex**](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare/e2325c81...master#3cc0701a2668b990684d9d4f9f430f46f63b24aa)

[View file @7fca65b](https://spotlite.nih.gov/ncats/scaffoldanalytics/blob/7fca65b4814833372e08489c48859581ecb70925/SI/SI.tex)

This diff is collapsed. Click to expand it.

[**fig/details\_all3\_noSNG.png**](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare/e2325c81...master#0345846e964f8b7e6a214f1e7bd93bf33a6c44dc) 0 → 100644

[View file @7fca65b](https://spotlite.nih.gov/ncats/scaffoldanalytics/blob/7fca65b4814833372e08489c48859581ecb70925/fig/details_all3_noSNG.png)

77.9 KB

[**fig/mol3\_RGtool\_scafpie\_iter.png**](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare/e2325c81...master#ab4625bce4b93b4b12687b8afec84507b7293842)

[View file @7fca65b](https://spotlite.nih.gov/ncats/scaffoldanalytics/blob/7fca65b4814833372e08489c48859581ecb70925/fig/mol3_RGtool_scafpie_iter.png)

236 KB | **W:** 933px | **H:** 572px

344 KB | **W:** 1201px | **H:** 757px

[**fig/spotviz\_scafpie\_tooltip.png**](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare/e2325c81...master#7d34ccbddd8d9da7ac881d5d6cafc53bd75e87ea)

[View file @7fca65b](https://spotlite.nih.gov/ncats/scaffoldanalytics/blob/7fca65b4814833372e08489c48859581ecb70925/fig/spotviz_scafpie_tooltip.png)

* 2-up

* Swipe

* Onion skin

[**fig/statcompare\_frames\_RGtool\_transparent\_density.png**](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare/e2325c81...master#5aac9b215d9632a4495c236e34034b6cf95c173a) 0 → 100644

[View file @7fca65b](https://spotlite.nih.gov/ncats/scaffoldanalytics/blob/7fca65b4814833372e08489c48859581ecb70925/fig/statcompare_frames_RGtool_transparent_density.png)

727 KB

[**pres/Backup and References - RGroup Stat stuff not presented at ACS.pptx**](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare/e2325c81...master#ea92a29f451471730446eb8bdd330c07a62584be)

[View file @7fca65b](https://spotlite.nih.gov/ncats/scaffoldanalytics/blob/7fca65b4814833372e08489c48859581ecb70925/pres/Backup%20and%20References%20-%20RGroup%20Stat%20stuff%20not%20presented%20at%20ACS.pptx)

No preview for this file type

[**scaffoldAnalyticsJCIM.pdf**](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare/e2325c81...master#49274ce5aa17b2cfa8d5b0dee94a0f2986b8fa6b) 0 → 100644

[View file @7fca65b](https://spotlite.nih.gov/ncats/scaffoldanalytics/blob/7fca65b4814833372e08489c48859581ecb70925/scaffoldAnalyticsJCIM.pdf)

No preview for this file type

[**scaffoldAnalyticsJCIM.tex**](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare/e2325c81...master#57eacaa2478e9f8de477650df3c4a120b229c96b)

[View file @7fca65b](https://spotlite.nih.gov/ncats/scaffoldanalytics/blob/7fca65b4814833372e08489c48859581ecb70925/scaffoldAnalyticsJCIM.tex)

|  |  |  |
| --- | --- | --- |
| ... | ... | @@ -64,21 +64,24 @@ |
|  |  | *%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%* |
|  |  | **\author**{Deepak Bandyopadhyay} |
|  |  | **\email**{debug22@gmail.com} |
|  |  | **\altaffiliation**{Current address: Janssen Pharmaceutical Companies of Johnson and Johnson, McKean and Welsh Roads, Spring House, PA 19477} |
|  |  | **\author**{Constantine Kreatsoulas} |
|  |  | **\author**{Pat G. Brady} |
|  |  | **\author**{Joseph Boyer} |
|  |  | **\author**{Zangdong He} |
|  |  | **\author**{Genaro Scavello Jr.} |
|  |  | **\affiliation**[GSK]{GlaxoSmithKline, 1250 S. Collegeville Rd, Collegeville, PA 19426} |
|  |  | **\author**{Dac-Trung Nguyen} |
|  |  | **\author**{Tyler Peryea} |
|  |  | **\author**{Ajit Jadhav} |
|  |  | **\author**{Dac-Trung Nguyen} |
|  |  | **\email**{nguyenda@mail.nih.gov} |
|  |  | **\author**{Rajarshi Guha} |
|  |  | **\email**{guhar@mail.nih.gov} |
|  |  | **\author**{Ajit Jadhav} |
|  |  | **\email**{ajadhav@mail.nih.gov} |
|  |  | *%\phone{+123 (0)123 4445556}* |
|  |  | *%\fax{+123 (0)123 4445557}* |
|  |  | **\affiliation**[NCATS]{National Center for Advancing Translational Science, 9800 Medical Center Drive, Rockville, MD 20850} |
|  |  | **\affiliation**[NCATS]{National Center for Advancing Translational |
|  |  | Science, 9800 Medical Center Drive, Rockville, MD 20850} |
|  |  | **\altaffiliation**{Current address: Vertex Pharmaceuticals, 50 Northern Avenue, Boston MA 02210} |
|  |  |  |
|  |  | *%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%* |
|  |  | *%% The document title should be given as usual. Some journals require* |
| ... | ... | @@ -133,32 +136,29 @@ |
|  |  | *%% if an abstract is not used by the target journal.* |
|  |  | *%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%* |
|  |  | \begin{abstract} |
|  |  | We present a method for visualizing and navigating large and diverse |
|  |  | chemical spaces, such as screening datasets, along with their |
|  |  | activities and properties. Our approach is to annotate the data with |
|  |  | all possible scaffolds contained within each molecule using an |
|  |  | exhaustive algorithm developed at NCATS. We have developed a |
|  |  | Spotfire visualization, coupled to a fuzzy clustering approach, that |
|  |  | is used to drive the hit triage process. Progression decisions can |
|  |  | be made using aggregate scaffold parameters and data from multiple |
|  |  | datasets merged at the scaffold level. This visualization easily |
|  |  | reveals overlaps that help prioritize hits, highlight tractable |
|  |  | series and posit ways to combine aspects of multiple hits. The SAR |
|  |  | of a large and complex hit is automatically mapped into all |
|  |  | constituent scaffolds making it possible to navigate, via any shared |
|  |  | scaffold, to all related hits. This scaffold ``walking'' helps |
|  |  | address bias toward a handful of potent and ligand-efficient |
|  |  | molecules at the expense of coverage of chemical space. We compare |
|  |  | the NCATS scaffold generation method with published screening triage |
|  |  | methods such as nearest-neighbor clustering, data-driven clustering |
|  |  | and scaffold networks. Similarities and differences in the methods |
|  |  | are evaluated qualitatively and statistically. The workflow of a |
|  |  | Spotfire visualization used in combination with fuzzy clustering and |
|  |  | structure annotation provides a novel view of large and diverse |
|  |  | datasets. This allows teams to effortlessly navigate between |
|  |  | structurally related molecules and enriches the population of leads |
|  |  | considered and progressed in a manner complementary to established |
|  |  | approaches. |
|  |  | We present a method for visualizing and navigating large screening |
|  |  | datasets, taking into account their activities and properties. Our |
|  |  | approach is to annotate the data with all possible scaffolds |
|  |  | contained within each molecule. We have developed a Spotfire |
|  |  | visualization, coupled to a fuzzy clustering approach based on the |
|  |  | scaffold decomposition of the screening deck, that is used to drive |
|  |  | the hit triage process. Progression decisions can be made using |
|  |  | aggregate scaffold parameters and data from multiple datasets merged |
|  |  | at the scaffold level. This visualization easily reveals overlaps |
|  |  | that help prioritize hits, highlight tractable series and posit ways |
|  |  | to combine aspects of multiple hits. The SAR of a large and complex |
|  |  | hit is automatically mapped onto all constituent scaffolds making it |
|  |  | possible to navigate, via any shared scaffold, to all related hits. |
|  |  | This scaffold ``walking'' helps address bias toward a handful of |
|  |  | potent and ligand-efficient molecules at the expense of coverage of |
|  |  | chemical space. We considered two scaffold generation methods and |
|  |  | explored their similarities and differences both qualitatively and |
|  |  | quantitatively. The workflow of a Spotfire visualization used in |
|  |  | combination with fuzzy clustering and structure annotation provides |
|  |  | a intuitive view of large and diverse screening datasets. This |
|  |  | allows teams to effortlessly navigate between structurally related |
|  |  | molecules and enriches the population of leads considered and |
|  |  | progressed in a manner complementary to established approaches. |
|  |  | \end{abstract} |
|  |  |  |
|  |  | *%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%* |
| ... | ... | @@ -183,12 +183,15 @@ removal of promiscuous or otherwise undesirable chemotypes |
|  |  | **\cite**{Dahlin:2014fp}. See **\citeauthor**{Shun:2011sy} and |
|  |  | **\citeauthor**{Langer:2009mw} for a review of HTS triage approaches. |
|  |  |  |
|  |  | A key challenges in the triage step is to identify structure-activity |
|  |  | series - sets of compounds with similar or analogous structures that exhibit a |
|  |  | spectrum of activity, including lack of activity. Identifying such subsets allows one to have some |
|  |  | confidence in the presence of a structure-activity relationship amongst the |
|  |  | active compounds which enables a more efficient exploration of the chemical |
|  |  | space around the selected hits. A good review of computational methods to extract SAR from screening datasets can be found in **\citeauthor**{Wawer2010review}. |
|  |  | A key challenge in the triage step is to identify structure-activity |
|  |  | series - sets of compounds with similar or analogous structures that |
|  |  | exhibit a spectrum of activity (including lack of |
|  |  | activity). Identifying such subsets allows one to have some confidence |
|  |  | in the presence of a structure-activity relationship amongst the |
|  |  | active compounds which enables a more efficient exploration of the |
|  |  | chemical space around the selected hits. A good review of |
|  |  | computational methods to extract SAR from screening datasets can be |
|  |  | found in **\citeauthor**{Wawer2010review}. |
|  |  |  |
|  |  | Given that a SAR series is, ideally, defined in terms of a core structure along |
|  |  | with various decorations, the first step in the triage process is to |
| ... | ... | @@ -208,48 +211,46 @@ scaffold-activity landscape of a screening collection using a fuzzy |
|  |  | clustering method to group compounds based on scaffold membership. The |
|  |  | workflow includes methods to visualize the activity landscape as well |
|  |  | as methods to explore different regions of chemical space via shared |
|  |  | scaffolds. We also highlight the efficiency gains obtained by using |
|  |  | pre-computed scaffold associations rather than performing substructure |
|  |  | searches on the fly. |
|  |  | scaffolds. |
|  |  |  |
|  |  | **\subsection**{Related Work} |
|  |  | While there are many ways to generate a set of scaffolds from a |
|  |  | compound collection, a key step is to identify a |
|  |  | relevant subset of them or else aggregate them in a way that leads to a useful |
|  |  | clustering of active and inactive compounds. While the term ``useful'' is rather |
|  |  | subjective, it is easy to identify cases that are **\emph**{not} useful. Thus, 5- or |
|  |  | 6-member undecorated rings are likely not useful since they will occur in the |
|  |  | majority of compounds in a screening collection. At the other extreme, large, |
|  |  | extended scaffolds that are associated with very few compounds are also likely |
|  |  | not useful. |
|  |  |  |
|  |  | As a result, many approaches to scaffold aggregation have been described. A |
|  |  | natural approach is to consider a hierarchical aggregation. The Scaffold |
|  |  | Tree**\cite**{Ertl2011ScaffoldTree} and Scaffold Network**\cite**{Varin2011ScafNet} |
|  |  | define a hierarchical decomposition from more specialized larger scaffolds to |
|  |  | more inclusive smaller scaffolds. While the Scaffold Tree splits each larger |
|  |  | scaffold in exactly one way into two scaffolds with fewer rings, the Scaffold |
|  |  | Network performs an exhaustive decomposition into all possible smaller scaffolds |
|  |  | with fewer rings. Since some subscaffolds are shared with neighboring |
|  |  | scaffolds, this produces a network or graph rather than a tree. The Scaffold |
|  |  | Network Generator program**\cite**{Matlock2013SNG} is a fast, easy-to-use public |
|  |  | command line implementation of Scaffold Networks that can scale to millions of |
|  |  | compounds via aggregation of scaffolds from split datasets and parallel |
|  |  | execution. **\citeauthor**{Harper2004DDclus} use exhaustive enumeration to find all |
|  |  | Bemis-Murcko like frameworks in each molecule, and then recursively retain those |
|  |  | frameworks with highest aggregate activity and remove molecules that contain |
|  |  | them until a threshold is met, yielding a set of disjoint frameworks. Other |
|  |  | methods have used multiple common substructure (MCS), first proposed for finding |
|  |  | protein structural similarity**\cite**{Koch1997MCSprot}, for example |
|  |  | **\citeauthor**{Quintus2009MCS} and the ChemAxon product LibraryMCS. |
|  |  |  |
|  |  | To overcome the effect of small variations in heteroatoms (eg. O to S) mapping |
|  |  | otherwise similar molecules to different scaffolds, generalized or consolidated |
|  |  | scaffold representations have been proposed, such as the Reduced |
|  |  | Graph**\cite**{Barker2003RG}, the Bemis-Murcko scaffold**\cite**{BemisMurcko1996} and |
|  |  | the topological or 2D pharmacophore**\cite**{Schneider1999ScafHopTP}. Bemis-Murcko |
|  |  | like frameworks **\cite**{Harper2004DDclus} are generalized Bemis-Murcko scaffolds |
|  |  | where atom types and/or bond orders may be retained. |
|  |  | compound collection, a key step is to identify a relevant subset of |
|  |  | them or else aggregate them in a way that leads to a useful clustering |
|  |  | of active and inactive compounds. While the term ``useful'' is rather |
|  |  | subjective, it is easy to identify cases that are not useful. Thus, 5- |
|  |  | or 6-member undecorated rings are likely not useful since they will |
|  |  | occur in the majority of compounds in a screening collection. At the |
|  |  | other extreme, large, extended scaffolds that are associated with very |
|  |  | few compounds are also likely not useful. |
|  |  |  |
|  |  | As a result, many approaches to scaffold aggregation have been |
|  |  | described. A natural approach is to consider a hierarchical |
|  |  | aggregation. The Scaffold Tree**\cite**{Ertl2011ScaffoldTree} and Scaffold |
|  |  | Network**\cite**{Varin2011ScafNet} define a hierarchical decomposition |
|  |  | from more specialized larger scaffolds to more inclusive smaller |
|  |  | scaffolds. While the Scaffold Tree splits each larger scaffold in |
|  |  | exactly one way into two scaffolds with fewer rings, the Scaffold |
|  |  | Network performs an exhaustive decomposition into all possible smaller |
|  |  | scaffolds with fewer rings. Since some subscaffolds are shared with |
|  |  | neighboring scaffolds, this produces a network or graph rather than a |
|  |  | tree. **\citeauthor**{Harper2004DDclus} use exhaustive enumeration to |
|  |  | find all Bemis-Murcko like frameworks in each molecule, and then |
|  |  | recursively retain those frameworks with highest aggregate activity |
|  |  | and remove molecules that contain them until a threshold is met, |
|  |  | yielding a set of disjoint frameworks. Other methods have used |
|  |  | multiple common substructure (MCS), first proposed for finding protein |
|  |  | structural similarity**\cite**{Koch1997MCSprot}, for example |
|  |  | **\citeauthor**{Quintus2009MCS} and the ChemAxon product LibraryMCS. |
|  |  |  |
|  |  | *% To overcome the effect of small variations in heteroatoms (eg. O to S)* |
|  |  | *% mapping otherwise similar molecules to different scaffolds,* |
|  |  | *% generalized or consolidated scaffold representations have been* |
|  |  | *% proposed, such as the Reduced Graph\cite{Barker2003RG}, the* |
|  |  | *% Bemis-Murcko scaffold\cite{BemisMurcko1996} and the topological or 2D* |
|  |  | *% pharmacophore\cite{Schneider1999ScafHopTP}. Bemis-Murcko like* |
|  |  | *% frameworks \cite{Harper2004DDclus} are generalized Bemis-Murcko* |
|  |  | *% scaffolds where atom types and/or bond orders may be retained.* |
|  |  |  |
|  |  | Multiple scaffolds if present in a dataset can be inferred from the |
|  |  | scaffold tree decomposition**\cite**{ClarkLabute2008SAReport}. However in |
| ... | ... | @@ -277,13 +278,13 @@ that a molecule could well be assigned simultaneously to multiple |
|  |  | clusters, depending on the structural features present. Traditional |
|  |  | hard clustering methods will not allow this and as a result will tend |
|  |  | to assign such molecules to their own cluster, thus identifying them |
|  |  | as singletons, Such a result can be observed in real chemical datasets |
|  |  | as singletons. Such a result can be observed in real chemical datasets |
|  |  | as well -- for example, the complete-linkage clustering of the TCAMS |
|  |  | dataset**\cite**{Gamo2010,Calderon2011} has nearly 25**\%** of the 2000 |
|  |  | clusters containing just one compound, as shown in **\fref**{platypus}. |
|  |  |  |
|  |  | In contrast to these partitioning clusters, the methods described |
|  |  | herein rely on fuzzy clustering, where a molecule may be assigned to |
|  |  | here rely on fuzzy clustering, where a molecule may be assigned to |
|  |  | multiple clusters. Fuzzy clusters have been rarely used in |
|  |  | cheminformatics (for example **\cite**{Holliday2004,Richmond2013Galois}), |
|  |  | perhaps because they are hard to visualize and navigate. In this work |
| ... | ... | @@ -296,7 +297,7 @@ clustering of multiple overlapping scaffolds per molecule. |
|  |  | The datasets used to illustrate and visualize our methods were picked |
|  |  | to represent the kinds of screening datasets we expect the method to |
|  |  | be used on in practice. For example, the TCAMS dataset**\cite**{Gamo2010} |
|  |  | consists of 13.5k diverse hits from an antimalarial screen at GSK, |
|  |  | consists of 13,500 diverse hits from an antimalarial screen at GSK, |
|  |  | along with pIC50 against a susceptible strain of the malarial parasite |
|  |  | (3D7), percentage inhibition against a resistant strain (DD2), Hep G2 |
|  |  | hepatotoxicity, a few physical chemical properties (**\eg** molecular |
| ... | ... | @@ -304,7 +305,7 @@ weight, aromatic ring count, cLogP), and Inhibition Frequency Index |
|  |  | (IFI, a measure of promiscuity defined as the percentage of screens in |
|  |  | which a molecule inhibits over 50**\%**, **\cite**{Chakravorty2013IFI}). |
|  |  |  |
|  |  | The in-house kinase dataset shown (``Kinase X''), by contrast, is perhaps less |
|  |  | The in-house GSK kinase dataset shown (``Kinase X''), by contrast, is perhaps less |
|  |  | diverse but chosen to illustrate the power of this approach in joining |
|  |  | and merging datasets from multiple screens, combining their SAR to |
|  |  | design hybrid molecules, and making inferences about unknown activity |
| ... | ... | @@ -378,11 +379,18 @@ were generated from the molecule shown in \fref{scafmethod}(a) from |
|  |  | the TCAMS dataset. |
|  |  |  |
|  |  | \begin{figure} |
|  |  | (a)**\includegraphics**[width=2.5in]{fig/tcam1\_mol.png}**\\** |
|  |  | (b)**\includegraphics**[width=4in]{fig/tcam1\_RGscaf.png}**\\** |
|  |  | (c)**\includegraphics**[width=4in]{fig/tcam1\_GSKframes.png}**\\** |
|  |  | (d)**\includegraphics**[width=3in]{fig/tcam1\_SNG3.png} |
|  |  | **\caption**{Scaffold Decompositions for a molecule (a) from the TCAMS dataset with PubChem Compound ID 536182. (b) 5 scaffolds from NCATS R-Group Tool; (c) 24 GSK frameworks -- first 21 Bemis-Murcko Like and last 3 in the bottom row RECAP; (d) Scaffold Network generated by SNG, starting from top-level scaffold with four rings down to to all subscaffolds with two rings.} |
|  |  | *%(a)\includegraphics[width=2.5in]{fig/tcam1\_mol.png}\\* |
|  |  | *%(b)\includegraphics[width=4in]{fig/tcam1\_RGscaf.png}\\* |
|  |  | *%(c)\includegraphics[width=4in]{fig/tcam1\_GSKframes.png}\\* |
|  |  | *%(d)\includegraphics[width=3in]{fig/tcam1\_SNG3.png}* |
|  |  | *%\caption{Scaffold Decompositions for a molecule (a) from the TCAMS dataset with PubChem Compound ID 536182. (b) 5 scaffolds from NCATS R-Group Tool; (c) 24 GSK frameworks -- first 21 Bemis-Murcko Like and last 3 in the bottom row RECAP. (d) Scaffold Network generated by SNG, starting from top-level scaffold with four rings down to to all subscaffolds with two rings.}* |
|  |  |  |
|  |  | (a)**\includegraphics**[width=3in]{fig/tcam1\_mol.png}**\\** |
|  |  | **\vspace**{0.1in} |
|  |  | (b)**\includegraphics**[width=5in]{fig/tcam1\_RGscaf.png}**\\** |
|  |  | **\vspace**{0.1in} |
|  |  | (c)**\includegraphics**[width=5in]{fig/tcam1\_GSKframes.png} |
|  |  | **\caption**{Scaffold Decompositions for a molecule (a) from the TCAMS dataset with PubChem Compound ID 536182. (b) 5 scaffolds from NCATS R-Group Tool; (c) 24 GSK frameworks -- first 21 Bemis-Murcko Like and last 3 in the bottom row RECAP.} |
|  |  | **\label**{fig:scafmethod} |
|  |  | \end{figure} |
|  |  |  |
| ... | ... | @@ -414,11 +422,13 @@ Molecular Frameworks is subtly different, and its use has been |
|  |  | described in **\citeauthor**{Harper2004DDclus}. The two types of framework |
|  |  | we retain for this study are Bemis-Murcko-like**\cite**{BemisMurcko1996} |
|  |  | (with atom and bond orders retained) and |
|  |  | RECAP**\cite**{Lewell:1998aa}. Other fragmentation methods described by |
|  |  | **\citeauthor**{Harper2004DDclus} such as reduced graphs and classic |
|  |  | Bemis-Murcko scaffolds (without atom types and bond orders) were |
|  |  | skipped for the purposes of comparison; however there is no reason |
|  |  | these could not be included in the technique we are proposing. |
|  |  | RECAP**\cite**{Lewell:1998aa}. |
|  |  |  |
|  |  | *% Other fragmentation methods described by* |
|  |  | *% \citeauthor{Harper2004DDclus} such as reduced graphs and classic* |
|  |  | *% Bemis-Murcko scaffolds (without atom types and bond orders) were* |
|  |  | *% skipped for the purposes of comparison; however there is no reason* |
|  |  | *% these could not be included in the technique we are proposing.* |
|  |  |  |
|  |  | The input for the GSK frameworks code is a comma separated text file |
|  |  | with molecules encoded in a SMILES field. The code was modified by |
| ... | ... | @@ -426,10 +436,10 @@ adding scripts to export the fuzzy clusters in a tabular format rather |
|  |  | than prioritize them into mutually exclusive scaffolds as in |
|  |  | **\citeauthor**{Harper2004DDclus}. This step produces a file similar to |
|  |  | the R-group decomposition format described for the NCATS R-group tool |
|  |  | in **\sref**{rgtool}: framework ID, framework SMILES, molecule ID, SMILES |
|  |  | and replicated property/activity columns. There are no R-group |
|  |  | columns simply because this is not a default computation in the GSK |
|  |  | frameworks code. |
|  |  | in **\sref**{rgtool}, including the following key columns: **\emph**{framework |
|  |  | ID, framework SMILES, molecule ID, SMILES} and |
|  |  | **\emph**{properties/activities}. There are no R-group columns simply |
|  |  | because this is not a default computation in the GSK frameworks code. |
|  |  |  |
|  |  | The GSK frameworks found within the same molecule from TCAMS are shown |
|  |  | in **\fref**{scafmethod}(c). The reader will observe several differences |
| ... | ... | @@ -437,38 +447,38 @@ from the R-group tool: there are more scaffolds found, some clipped in |
|  |  | the middle of a linker rather than at a ring, and some redundancy |
|  |  | between multiple scaffolds. Also the current implementation does not |
|  |  | convert or unify tautomers among scaffolds, again leading to larger |
|  |  | numbers of scaffolds. We will see in Section **\ref**{sec:results} that |
|  |  | comprehensive coverage of fragments within each molecule can be both |
|  |  | good and bad. |
|  |  | numbers of scaffolds. *% We will see in Section \ref{sec:results} that* |
|  |  | *%comprehensive coverage of fragments within each molecule can be both* |
|  |  | *%good and bad.* |
|  |  |  |
|  |  | Further details on how to set up and run the GSK frameworks code are |
|  |  | provided in the Supplementary Material Section S4. *%\textbf{\textcolor{red}{TODO}}* |
|  |  |  |
|  |  | **\subsection**{Fragmentation Method: Scaffold Network Generator} |
|  |  | **\label**{sec:SNG} |
|  |  | The Scaffold Network Generator (SNG) **\cite**{Matlock2013SNG} is a |
|  |  | parallelizable and robust code to generate a hierarchical Scaffold |
|  |  | Network from any chemical dataset. The operation of this tool is |
|  |  | described on the web at |
|  |  | **\url**{https://bitbucket.org/swamidass/scaffold-network-generator/wiki/Home}. SNG |
|  |  | takes as input either a SMILES or an MDL SD-file, and we specify |
|  |  | options to generate the Network and ID Map files as two tabular |
|  |  | outputs. |
|  |  |  |
|  |  | The Network file lists each Scaffold with numeric ID, SMILES, Number |
|  |  | of Rings (which serves as the level in the hierarchy) and Subscaffolds |
|  |  | presented as a comma-separated numeric list. This list was converted into |
|  |  | a multi-line format, one for each subscaffold, as described in the |
|  |  | Supplementary Material in Section S5. |
|  |  |  |
|  |  | The ID Map file has two columns, mapping a Molecule ID from the |
|  |  | primary dataset to the Top-Level Scaffold (i.e. Murcko scaffold) |
|  |  | obtained by stripping all pendant groups but no rings. Using the ID |
|  |  | Map file followed by multiple iterations of the Network file one can |
|  |  | elucidate the entire Scaffold Network starting from each query |
|  |  | molecule, as described in a subsequent section. As an example the |
|  |  | scaffold network generated for a molecule from TCAMS is shown in |
|  |  | **\fref**{scafmethod}(d), as for the previous two methods. |
|  |  | *%\subsection{Fragmentation Method: Scaffold Network Generator}* |
|  |  | *%\label{sec:SNG}* |
|  |  | *%The Scaffold Network Generator (SNG) \cite{Matlock2013SNG} is a* |
|  |  | *%parallelizable and robust code to generate a hierarchical Scaffold* |
|  |  | *%Network from any chemical dataset. The operation of this tool is* |
|  |  | *%described on the web at* |
|  |  | *%\url{https://bitbucket.org/swamidass/scaffold-network-generator/wiki/Home}. SNG* |
|  |  | *%takes as input either a SMILES or an MDL SD-file, and we specify* |
|  |  | *%options to generate the Network and ID Map files as two tabular* |
|  |  | *%outputs.* |
|  |  | *%* |
|  |  | *%The Network file lists each Scaffold with numeric ID, SMILES, Number* |
|  |  | *%of Rings (which serves as the level in the hierarchy) and Subscaffolds* |
|  |  | *%presented as a comma-separated numeric list. This list was converted into* |
|  |  | *%a multi-line format, one for each subscaffold, as described in the* |
|  |  | *%Supplementary Material in Section S5.* |
|  |  | *%* |
|  |  | *%The ID Map file has two columns, mapping a Molecule ID from the* |
|  |  | *%primary dataset to the Top-Level Scaffold (i.e. Murcko scaffold)* |
|  |  | *%obtained by stripping all pendant groups but no rings. Using the ID* |
|  |  | *%Map file followed by multiple iterations of the Network file one can* |
|  |  | *%elucidate the entire Scaffold Network starting from each query* |
|  |  | *%molecule, as described in a subsequent section. As an example the* |
|  |  | *%scaffold network generated for a molecule from TCAMS is shown in* |
|  |  | *%\fref{scafmethod}(d), as for the previous two methods.* |
|  |  |  |
|  |  |  |
|  |  | **\section**{Methods: Data Integration and Visualization in Spotfire} |
| ... | ... | @@ -476,12 +486,13 @@ scaffold network generated for a molecule from TCAMS is shown in |
|  |  |  |
|  |  | Next, we describe how tabular scaffold output generated using the |
|  |  | NCATS R-group tool and other comparable methods is integrated into |
|  |  | Spotfire, our visualization tool of choice at GSK. |
|  |  | Spotfire, our visualization tool of choice at GSK. |
|  |  |  |
|  |  | **\subsection**{Data Table Generation and Linking} |
|  |  |  |
|  |  | \begin{figure} |
|  |  | **\includegraphics**[width=6in]{fig/details\_all2.png} |
|  |  | *%\includegraphics[width=6in]{fig/details\_all2.png}* |
|  |  | **\includegraphics**[width=6in]{fig/details\_all3\_noSNG.png} |
|  |  | **\caption**{Detailed schematic on how the output from clustering and |
|  |  | fragmentation methods are set up as data tables and linked together |
|  |  | with the main dataset in Spofire. Right inset: schematic color-coded |
| ... | ... | @@ -491,7 +502,7 @@ Spotfire, our visualization tool of choice at GSK. |
|  |  | \end{figure} |
|  |  |  |
|  |  | **\fref**{detaildevil} shows how the data tables output by the scaffold |
|  |  | generation methods considered here, are layered onto the primary data |
|  |  | generation methods considered here are layered onto the primary data |
|  |  | table in Spotfire. This primary data table is usually a direct import |
|  |  | of tabular molecule and activity data generated at GSK or available |
|  |  | from public datasets. What gets added is by and large similar, |
| ... | ... | @@ -515,13 +526,13 @@ following elements: |
|  |  | selectivity, ligand efficiency and molecular properties may be |
|  |  | highlighted on the X, Y, shape, size and color axes on a scatter |
|  |  | plot. This window is not illustrated in the interests of space, and |
|  |  | since the reader may already be familiar. |
|  |  | since the reader familiar with screening datasets may already have seen a similar view. |
|  |  | *%The main window is illustrated for the TCAMS dataset in \fref{spotviz}(a).* |
|  |  | **\item** {**\bf** Related Molecules Tab}: The purpose of this tab is to |
|  |  | implement the Scaffold Walking navigation described briefly earlier. |
|  |  | The setup is described for the NCATS R-group Tool decomposition, |
|  |  | though this tab applies to and can be set up analogously for any of |
|  |  | the other decompositions. The tab consists of two visualizations, |
|  |  | though this tab applies to and can be set up analogously for any |
|  |  | other decomposition. The tab consists of two visualizations, |
|  |  | illustrated for the TCAMS dataset in **\fref**{spotviz}(a): **\subitem** The |
|  |  | first one is a miniature version of the Main window, allowing the |
|  |  | user to select (in Spotfire, mark) molecules of interest without |
| ... | ... | @@ -551,7 +562,7 @@ following elements: |
|  |  | important or unimportant for activity. |
|  |  |  |
|  |  | **\item** {**\bf** Scaffolds and R-groups Tab}: This tab, currently specific to the NCATS R-group tool method for generating scaffolds, contains two visualizations, as illustrated for the TCAMS dataset in **\fref**{spotviz}(b)--(c): |
|  |  | **\subitem** The first is a scatter plot display of the Scaffolds table, displaying scaffolds Complexity and Count and aggregate activity of each scaffold either on the axes or using the Size and Shape dimensions. Here scaffolds of lesser interest (for example with low complexity or count) can be identified and tagged to remove them from the analysis. Conversely, scaffolds of high interest, for example with many active members or high aggregate ligand efficiency, may be tagged into separate categories. |
|  |  | **\subitem** The first is a scatter plot display of the Scaffolds table, displaying scaffold Complexity and Count and aggregate activity of each scaffold either on the axes or using the Size and Shape dimensions. Here scaffolds of lesser interest (for example with low complexity or count) can be identified and tagged to remove them from the analysis. Conversely, scaffolds of high interest, for example with many active members or high aggregate ligand efficiency, may be tagged into separate categories. |
|  |  | **\subitem** The second plot is an R-group table, **\ie** a Table view of the {RG}decomp table limited to data records that have been marked, **\ie** molecules that lie in scaffolds currently marked. The table is sorted first by scaffold and then by primary activity, and molecular fields such as Scaffold SMILES, Molecule SMILES and R-groups $R\\_1..R\\_n$ are rendered using an appropriate depiction package - at GSK this is {JChem}. This table may be exported to Excel as an on-the-fly R-group table of the scaffolds of interest. |
|  |  |  |
|  |  | \end{itemize} |
| ... | ... | @@ -603,7 +614,7 @@ methods we can evaluate the overall similarity of the two |
|  |  | fragmentation schemes, and also independently score the usefulness of |
|  |  | $A$ and $B$ to connect compound(s) of interest to related molecules. |
|  |  |  |
|  |  | For any compound $C$ and ontology $A$, define the structure group of |
|  |  | For any compound $C$ and fragmentation $A$, define the structure group of |
|  |  | $C$ under $A$, $C\_A$ as the set of compounds that share fragments from |
|  |  | $A$ with compound $C$. Similarly define the structure group of $C$ |
|  |  | under $B$, $C\_B$ as the set of compounds that share fragments from |
| ... | ... | @@ -628,13 +639,27 @@ In constrast, the Proportion of Information Unique to $A, PIU\_A$ uses the set di |
|  |  | PIU\_A(C) = **\|** C\_A **\setminus** C\_B **\|** / **\|** C\_A **\cup** C\_B **\|** = 1 - PI\_c(B) |
|  |  | \end{equation} |
|  |  |  |
|  |  | When comparing one fragmentation method against another, we often see that one method utilizes a vastly greater number of shared fragments than the other in order to connect compound $C$ to a very similarly sized structure group. To capture this tendency and reward methods that connect molecules to related ones efficiently rather than exhaustively, we define a Fragment Efficiency measure as follows. Let $frag\_A**(**C**)**$ be the set of fragments of compound $C$ in ontology $A$ that connects $C$ to its structure group $C\_A$. Similarly define $frag\_B**(**C**)**$ for ontology $B$. Then: |
|  |  | When comparing one fragmentation method against another, we often see |
|  |  | that one method utilizes a larger number of shared fragments than the |
|  |  | other in order to connect compound $C$ to a very similarly sized |
|  |  | structure group. To capture this tendency and reward methods that |
|  |  | connect molecules to related ones efficiently rather than |
|  |  | exhaustively, we define a Fragment Efficiency measure as follows. Let |
|  |  | $frag\_A**(**C**)**$ be the set of fragments of compound $C$ in fragmentation |
|  |  | $A$ that connects $C$ to its structure group $C\_A$. Similarly define |
|  |  | $frag\_B**(**C**)**$ for fragmentation $B$. Then: |
|  |  | \begin{equation} |
|  |  | FragEff\_A(C) = **\|** C\_A **\|** / **\|** frag\_A(C) **\|** |
|  |  | \end{equation} |
|  |  |  |
|  |  |  |
|  |  | These statistics can be averaged over all compounds in a dataset to yield the Average Common Proportion (ACP), Average Proportion of Information (API), Average Proportion of Information Unique to A (APIU) or Average Fragment Efficiency (AFE). Other statistics can also be applied to the distribution of $CP$, $PI$ or $PIU$ to characterize the dataset and overlapping scaffolds used to characterize it. |
|  |  | These statistics can be averaged over all compounds in a dataset to |
|  |  | yield the Average Common Proportion (ACP), Average Proportion of |
|  |  | Information (API), Average Proportion of Information Unique to A |
|  |  | (APIU) or Average Fragment Efficiency (AFE). Other statistics can |
|  |  | also be applied to the distribution of $CP$, $PI$ or $PIU$ to |
|  |  | characterize the dataset and overlapping scaffolds used to |
|  |  | characterize it. |
|  |  |  |
|  |  | Our methods extend the similarity score of **\citeauthor**{Torres2009} to |
|  |  | overlapping clusters by using a set of clusters derived from them, |
| ... | ... | @@ -652,11 +677,18 @@ defined as follows: |
|  |  | Here we illustrate some of our key findings and use cases on a few datasets. |
|  |  |  |
|  |  | **\subsection**{Use Case: Scaffold Progression and Prioritization} |
|  |  | Aggregate statistics such as maximum, minimum, mean and standard deviation, computed at a per scaffold level, may be useful in prioritizing scaffolds. For example, **\fref**{RGTaggr} shows the six scaffolds contained within a tricyclic molecule from TCAMS (Molecule 2, CID: 533945) ranked by average activity and IFI. This ordering may be used to determine which substructures are most important for the molecule's activity, and use this information to design or test further compounds. |
|  |  | Aggregate statistics such as maximum, minimum, mean and standard |
|  |  | deviation, computed at a per scaffold level, may be useful in |
|  |  | prioritizing scaffolds. For example, **\fref**{RGTaggr} shows the six |
|  |  | scaffolds contained within a tricyclic molecule from TCAMS (Molecule |
|  |  | 2, CID: 533945) ranked by average activity and IFI. This ordering may |
|  |  | be used to determine which substructures are most important for the |
|  |  | molecule's activity, and use this information to design or test |
|  |  | further compounds. |
|  |  |  |
|  |  | \begin{figure} |
|  |  | **\includegraphics**[width=5in]{fig/RGT\_aggr\_prop2.png} |
|  |  | **\caption**{All **\~**5000 scaffolds from the TCAMS dataset ranked by average pIC50 in the P.~falciparum 3D7 strain and Inhibition Frequency Index. The six scaffolds contained in Molecule 2 (CID: 533945) are shown in increasing order of average } |
|  |  | **\caption**{All **\~**5000 scaffolds from the TCAMS dataset ranked by average pIC50 in the P.~falciparum 3D7 strain and Inhibition Frequency Index. The six scaffolds contained in Molecule 2 (CID: 533945) are shown in increasing order of average pIC50.} |
|  |  | **\label**{fig:RGTaggr} |
|  |  | \end{figure} |
|  |  |  |
| ... | ... | @@ -735,7 +767,7 @@ Another intriguing result is seen by observing a new tricyclic series that share |
|  |  | {**\bf** NCATS R-group tool}: As opposed to the clustering method, |
|  |  | if any two molecules share a common substructure that meets the standards required of a scaffold by the NCATS method (**\eg** being bordered by rings on each end), then those molecules will be found to contain that shared substructure as a scaffold and their activities will be used to compute aggregate properties for it. |
|  |  |  |
|  |  | {**\bf** Other Scaffold Generation Methods}: As shown earlier, even though other scaffold generation methods (GSK frameworks and Scaffold Network Generator) differed in their implementation details and produced different numbers of scaffolds for the same molecule, they were roughly equivalent in a qualitative sense with regard to the insights obtained. Due to substantial overlap between sets of scaffolds, ring systems responsible for activity of a molecule were generally revealed by any of the methods. For example, the insights mentioned in **\sref**{scafwalk} were more or less consistent across the three methods. However, there were cases where the Frameworks revealed negative information about a fragment being not important for activity that is also useful for a drug discovery scientist. For example, in **\fref**{frameswalk} a substructure is highlighted that is on the aggregate inactive and could be removed or substituted. This insight is not available from the other methods since they don't define or find that fragment as a scaffold. |
|  |  | {**\bf** Other Scaffold Generation Methods}: As shown earlier, even though other scaffold generation methods (represented here by GSK frameworks) differed in their implementation details and produced different numbers of scaffolds for the same molecule, they were roughly equivalent in a qualitative sense with regard to the insights obtained. Due to substantial overlap between sets of scaffolds, ring systems responsible for activity of a molecule were generally revealed by any of the methods. For example, the insights mentioned in **\sref**{scafwalk} were more or less consistent across the methods. However, there were cases where the Frameworks revealed negative information about a fragment being not important for activity that is also useful for a drug discovery scientist. For example, in **\fref**{frameswalk} a substructure is highlighted that is on the aggregate inactive and could be removed or substituted. This insight is not available from SSSR-based scaffolding methods such as the NCATS R-group tool since they don't define or find that fragment as a scaffold. |
|  |  |  |
|  |  | \begin{figure} |
|  |  | **\includegraphics**[width=5in]{fig/mol1\_frames\_scafpie.png} |
| ... | ... | @@ -743,15 +775,18 @@ if any two molecules share a common substructure that meets the standards requir |
|  |  | **\label**{fig:frameswalk} |
|  |  | \end{figure} |
|  |  |  |
|  |  | \begin{figure} |
|  |  | **\includegraphics**[width=5in]{fig/mol2\_SNG\_relmol\_trellis.png} |
|  |  | **\caption**{Using SNG to find Related Molecules a few hops away in the Scaffold Network that have a desirable property, in this case activity against the resistant DD2 strain of P.~falc. The star symbol shows the location of the parent molecule in this plot, and the compass device at the origin shows the direction of favorable properties (+X and +Y axes).} |
|  |  | **\label**{fig:SNGwalk} |
|  |  | \end{figure} |
|  |  | *%\begin{figure}* |
|  |  | *%\includegraphics[width=5in]{fig/mol2\_SNG\_relmol\_trellis.png}* |
|  |  | *%\caption{Using SNG to find Related Molecules a few hops away in the Scaffold Network that have a desirable property, in this case activity against the resistant DD2 strain of P.~falc. The star symbol shows the location of the parent molecule in this plot, and the compass device at the origin shows the direction of favorable properties (+X and +Y axes).}* |
|  |  | *%\label{fig:SNGwalk}* |
|  |  | *%\end{figure}* |
|  |  | *%* |
|  |  | *%Using the Scaffold Network to visualize Related Molecules, one is able to deconvolute scaffolds with different numbers of rings (which serves as the level in the hierarchy) by trellising on it, and thus gather together substructures of comparable size. As shown in \fref{SNGwalk}, with the same tricyclic molecule as in \fref{scafwalk2}, we use Scaffold Networks to solve a selectivity issue. Members of the original scaffold, shown in light brown, tend to have some activity against the susceptible 3D7 strain of $P.~falciparum$ but be inactive against the multidrug-resistant DD7 strain. Using the scaffold network to go one level down (to molecules with two rings) and then two levels up (up to 5 rings) we obtain many new molecules that are a few hops away from the top-level tricyclic scaffold while still sharing some aspect of it. Looking at where these new Related Molecules are placed, we mark one scaffold with 3 rings that has a quinazoline linked to a phenyl, and one with 4 rings that contains an indoloquinazoline (also discovered by scaffold walking with the NCATS R-group tool as described above) linked to a phenyl. These two scaffolds are highlighted since their members consistently have high activity against the resistant DD2 strain, and thus this would be a desirable scaffold hop (or walk) towards chemical space that is more useful in overcoming resistance from the malarial parasite.* |
|  |  |  |
|  |  | *%To summarize, all three multiple-scaffold decomposition methods considered in this study, \ie NCATS R-group Tool, GSK Frameworks and Scaffold Network Generator give comparable insights when exploring the TCAMS dataset, with some differences stemming from individual substructures that are considered shared scaffolds or not by the individual methods. We now explore these overlaps, similarities and differences in the aggregate using the statistical methods described earlier in \sref{statmethod}.* |
|  |  |  |
|  |  | Using the Scaffold Network to visualize Related Molecules, one is able to deconvolute scaffolds with different numbers of rings (which serves as the level in the hierarchy) by trellising on it, and thus gather together substructures of comparable size. As shown in **\fref**{SNGwalk}, with the same tricyclic molecule as in **\fref**{scafwalk2}, we use Scaffold Networks to solve a selectivity issue. Members of the original scaffold, shown in light brown, tend to have some activity against the susceptible 3D7 strain of $P.~falciparum$ but be inactive against the multidrug-resistant DD7 strain. Using the scaffold network to go one level down (to molecules with two rings) and then two levels up (up to 5 rings) we obtain many new molecules that are a few hops away from the top-level tricyclic scaffold while still sharing some aspect of it. Looking at where these new Related Molecules are placed, we mark one scaffold with 3 rings that has a quinazoline linked to a phenyl, and one with 4 rings that contains an indoloquinazoline (also discovered by scaffold walking with the NCATS R-group tool as described above) linked to a phenyl. These two scaffolds are highlighted since their members consistently have high activity against the resistant DD2 strain, and thus this would be a desirable scaffold hop (or walk) towards chemical space that is more useful in overcoming resistance from the malarial parasite. |
|  |  | To summarize, both multiple-scaffold decomposition methods considered in this study, **\ie** NCATS R-group Tool and GSK Frameworks give comparable insights when exploring the TCAMS dataset, with some differences stemming from individual substructures that are considered shared scaffolds or not by the individual methods. We now explore these overlaps, similarities and differences in the aggregate using the statistical methods described earlier in **\sref**{statmethod}. |
|  |  |  |
|  |  | To summarize, all three multiple-scaffold decomposition methods considered in this study, **\ie** NCATS R-group Tool, GSK Frameworks and Scaffold Network Generator give comparable insights when exploring the TCAMS dataset, with some differences stemming from individual substructures that are considered shared scaffolds or not by the individual methods. We now explore these overlaps, similarities and differences in the aggregate using the statistical methods described earlier in **\sref**{statmethod}. |
|  |  |  |
|  |  | **\subsection**{Statistical Comparison of Scaffold-Generation Methods}**\label**{sec:statcomp} |
|  |  |  |
| ... | ... | @@ -806,21 +841,22 @@ We compare the Common Proportion, Proportion of Information Unique (PIU) and Fra |
|  |  |  |
|  |  |  |
|  |  | \begin{figure} |
|  |  | **\includegraphics**[width=6in]{fig/statcompare\_frames\_RGtool.png} |
|  |  | **\caption**{Comparison of the Common Proportion, PIU and FragEff statistics (described in the text) between GSK Frameworks (method ``{**\bf** A}'') and NCATS R-group Tool (``{**\bf** B}'') for all molecules in TCAMS. PIU of A and B are on the X and Y axes, and the points are colored by the log ratio of FragEff(B) to FragEff(A) and sizes by Common Proportion.} |
|  |  | **\includegraphics**[width=6in]{fig/statcompare\_frames\_RGtool\_transparent\_density.png} |
|  |  | **\caption**{Comparison of the Common Proportion, PIU and FragEff statistics (described in the text) between GSK Frameworks (method ``{**\bf** A}'') and NCATS R-group Tool (``{**\bf** B}'') for all molecules in TCAMS. PIU of A and B are on the X and Y axes, and the points are colored by the log ratio of FragEff(B) to FragEff(A) and sized by Common Proportion.} |
|  |  | **\label**{fig:statcompare} |
|  |  | \end{figure} |
|  |  |  |
|  |  | We can make a few observations from this data and plot: |
|  |  | \begin{enumerate} |
|  |  | **\item** The two methods allow one to access different sets of molecules starting from any molecule in TCAMS - the average overlap in their coverage is 40**\%**. |
|  |  | **\item** On the average, one can link to about twice as many molecules with the GSK frameworks; however, this is because on the average there are 6 times more frameworks than NCATS scaffolds, so the fragment efficiency is actually 3 times greater for NCATS scaffolds. One could also argue that many of the framework-only links (**\eg** variously decorated benzene rings) are not useful. |
|  |  | **\item** The two methods allow one to access different sets of molecules starting from any molecule in TCAMS - the average overlap in their coverage is 40**\%**. Molecules with almost complete overlap in coverage have high Common Proportion and lie towards the bottom left of **\fref**{statcompare}. |
|  |  | **\item** On the average, frameworks connected to a compound add more unique information than NCATS scaffolds connected to the same compound - this is seen in the higher density near the X-axis in **\fref**{statcompare}. |
|  |  | **\item** On the average, one can link to about twice as many molecules with the GSK frameworks; however, this is because on the average there are 6 times more frameworks than NCATS scaffolds. *%, \ie NCATS scaffolds are three times more fragment-efficient for this dataset. %, so the fragment efficiency is actually 3 times greater for NCATS scaffolds. %One could also argue that many of the framework-only links (\eg variously decorated benzene rings) are not useful.* |
|  |  | **\item** The outliers are interesting. At one end, compounds in a rare tautomer are unified with the dominant one by the NCATS tool (high fragment efficiency), but left as singletons by the frameworks (low fragment efficiency). And compounds whose only link with other molecules would be a benzene ring or similar low complexity scaffold remain singletons with the R-group tool (lower fragment efficiency). |
|  |  | \end{enumerate} |
|  |  |  |
|  |  | **\subsection**{Statistical Basis of Structure-Activity Relationships (SAR)}**\label**{sec:statSAR} |
|  |  |  |
|  |  | We have also explored the Common Proportion measure to get at the overlap between structure and activity, **\ie** to investigate the statistical basis of SAR based on overlapping scaffolds. To do this, we define an Activity Group of compound $C$, $C\_{act}$ as a set of compounds with activity bordering that of $C$ (in this case pIC50 in the 3D7 antimalarial assay). The activity group of each compound $C$ is defined to be the same size as the structure group, denoted $C\_A$ or $C\_B$ above, and generalized here as $C\_{str}$. This is done arbitrarily for ease of statistical calculations, being cognizant that it may soimetimes lead to issues such as breaking a large list of compounds with the same activity arbitrarily, especially when picking a small list of activity neighbors for compounds with a small structure group. |
|  |  | We have also explored the Common Proportion measure to get at the overlap between structure and activity, **\ie** to investigate the statistical basis of SAR based on overlapping scaffolds. To do this, we define an Activity Group of compound $C$, $C\_{act}$ as a set of compounds with activity bordering that of $C$ (in this case pIC50 in the 3D7 antimalarial assay). The activity group of each compound $C$ is defined to be the same size as the structure group, denoted $C\_A$ or $C\_B$ above, and generalized here as $C\_{str}$. This is done arbitrarily for ease of statistical calculations, being cognizant that it may sometimes lead to issues such as breaking a large list of compounds with the same activity arbitrarily, especially when picking a small list of activity neighbors for compounds with a small structure group. |
|  |  |  |
|  |  | The structure-activity common proportion, $CP\_{str,act}**(**C**)**$ then measures the overlap between the structure neighbors of $C$, **\ie** compounds sharing any scaffold with $C$, and its activity neighbors, **\ie** compounds in a similar activity range. |
|  |  | \begin{equation} |
| ... | ... | @@ -852,104 +888,69 @@ On the other hand, there are also some compounds, highlighted on the left side o |
|  |  | **\section**{Discussion} |
|  |  | **\label**{sec:discussion} |
|  |  |  |
|  |  |  |
|  |  | \begin{itemize} |
|  |  | *%\item \sout{Discuss performance - scaffold generation is usually a one time procedure* |
|  |  | *% for a given screening deck. Furthermore, the scaffold generating process will* |
|  |  | *% associate compounds with scaffolds, so looking up scaffold membership is very* |
|  |  | *% fast. When considering scaffold similarity, usual performance bottlenecks* |
|  |  | *% occur, same as for other similarity applications. Can we include some* |
|  |  | *% performance numbers?}* |
|  |  | *%\item \sout{we haven't discussed removal of promiscuous compounds/chemotypes,* |
|  |  | *% undesirable chemotypes (ie PAINS etc), these are generally a separate and* |
|  |  | *% independent step from the scaffold identification process}* |
|  |  | **\item** Ranking scaffolds is a key step in prioritizing hits in a scaffold-based |
|  |  | approach. Still a subjective issue and many ways to do it. Not clear that |
|  |  | there is a single optimal way. Discussion points could include |
|  |  | \begin{itemize} |
|  |  | **\item** Size and complexity - depending on the application (and |
|  |  | dataset) smaller scaffolds maybe more relevant or useful than |
|  |  | larger scaffolds. Scaffold key approach **\cite**{Ertl:2014eu} |
|  |  | quantifies this aspect. In some scenarios, the dataset may consist |
|  |  | of small scaffolds in general (**\eg** fragment screening |
|  |  | libraries). Complexity may also play a role in the practicality of |
|  |  | using a scaffold - highly complex scaffolds may not be easily |
|  |  | synthesizable or purchasable. |
|  |  | **\item** Promiscuity of the scaffold (does it show up in many active |
|  |  | compounds across different assays?) itself or do the scaffold |
|  |  | members, in aggregate, show high degree of promiscuity? |
|  |  | **\item** Do the members of a scaffold represent a SAR (``SARability'')? |
|  |  | Could be characterized quantitatively, but such approaches would |
|  |  | be hampered by the size of the member set. Larger scaffolds will |
|  |  | tend to have fewer members. Alternatively, characterize the |
|  |  | activity landscape of the scaffold member set - overly smooth or |
|  |  | overly rough landscapes are not useful. But still tricky to decide |
|  |  | what a sufficiently smooth (or rough) landscape is. |
|  |  | **\item** Alternatively, in absence of SAR, does a given scaffold show |
|  |  | an enrichment of activity (or just actives) compared to other |
|  |  | scaffolds (after having taken parent-relationships in to account - |
|  |  | similar to GO enrichment analyses). |
|  |  | \end{itemize} |
|  |  | **\item** Privileged scaffolds - sometimes could go in looking for them, |
|  |  | but usually one identifies such privileged scaffolds in a |
|  |  | retrospective fashion, across multiple HTS campaigns. Relevance / |
|  |  | importance to scaffold based triage? |
|  |  | \end{itemize} |
|  |  |  |
|  |  | We observed in the Kinase X scaffold prioritization use case that aggregate statistics on overlapping scaffolds can be a powerful tool to find your best compounds and progress them based on the statistics of overlapping scaffolds rather than non-overlapping clusters. Spotfire can interactively be used to drill down from aggregate properties to individual compounds, so decisions are still made using real and individual compound data after using aggregate data as a filter. |
|  |  |  |
|  |  | We also observed from the Kinase X data fusion use case that an ELT dataset with no activity fields and virtual molecules can be combined with disparate other datasets. These datasets need not have common activity fields, and disparate activity fields if present may be normalized to a common scale. Such datasets can then be combined and merged at the scaffold level to derive insights such as: |
|  |  | \begin{itemize} |
|  |  | **\item** Are there regions of chemical space that are unique to one dataset and should be explored further to enrich the chemical diversity available? |
|  |  | **\item** Conversely, are there chemotypes that substantially overlap between the datasets? This might increase confidence in otherwise noisy screening data and provide a coherent SAR picture (including defining series) that might not be available from individual datasets. |
|  |  | **\item** Can active or otherwise interesting scaffolds from one or a few datasets point to latent or unmeasured actives in other datasets that can then be synthesized (if virtual) or measured (if available) to confirm their activity or add a SAR data point. |
|  |  | **\item** If there is a common chemical motif behind several actives, the method will reveal it automatically without relying on a chemist's having seen and flagged the similarity, even if the molecules concerned are disparate looking, few, and scattered across multiple datasets. |
|  |  | **\item** A cohesive view of actives, inactives, and SAR from a set of screening campaigns helps to summarize the effort and return to it if needed without relying on memory, lab notebooks or manual effort in collating and remining the data. |
|  |  | \end{itemize} |
|  |  |  |
|  |  | Anecdotal comparison of overlapping scaffolds to commonly used partition-based clustering methods such as complete linkage clustering shows that the latter methods may divide very similar molecules arbitrarily into two different clusters, after which the two sets of molecules remain distinct and manual compilation of sets of related clusters is necessary to see all molecules related to a single one. With overlapping scaffolds, if two molecules share a common substructure, these methods will find it -- provided of course that it is a legitimate scaffold for a particular method and thresholds have not been set on scaffold count or complexity that exclude small groups or simple substructures. |
|  |  |  |
|  |  | Our statistical methods for quantifying the similarity and differences |
|  |  | between different methods of computing ovrelapping scaffolds we hope |
|  |  | are a useful contribution in their own right. Common proportion, |
|  |  | proportion of information unique to a method, fragment efficiency and |
|  |  | normalized common proportion give practitioners simple and objective |
|  |  | ways to determine which fragmentation methods work best for meeting |
|  |  | the goals of chemists to find interesting active molecules related to |
|  |  | a hit. As mentioned earlier they also extend the applicability of |
|  |  | While threshold-based hit selection is a prevalent approach in the |
|  |  | analysis of high throughput screening datasets, it is simplistic and |
|  |  | ignores the extra information encoded in chemical structure. Thus |
|  |  | scaffold based analysis of high throughput screening datasets |
|  |  | represents a truly data-driven approach to hit triage that attempts to |
|  |  | make use of **\emph**{all} the data collected from a high throughput |
|  |  | screen. Ranking scaffolds is a key step in prioritizing hits in a |
|  |  | scaffold-based approach, and while there are many ways to generate a |
|  |  | ranking, it is not obvious which method is optimal (or whether there |
|  |  | is even a single, optimal method). |
|  |  |  |
|  |  | In this study we have presented some alternate approaches to scaffold |
|  |  | prioritization that make use of aggregate statistics based on |
|  |  | overlapping scaffolds, with the goal of provide a quantitative basis |
|  |  | for the comparison of different scaffold-based analysis schemes. |
|  |  |  |
|  |  | We observe that this approach, applied to the kinase ``X'' dataset can |
|  |  | be a powerful tool to identify promising hits. While simple to implement, |
|  |  | the use of the Spotfire workflow allows interactive drill |
|  |  | down from aggregate properties to individual compounds. In essence, |
|  |  | the workflow described here enables decisions on individual compounds, |
|  |  | but using the aggregated data as a filter. Another advantage of the |
|  |  | workflow described here is that it supports the inclusion of ``virtual |
|  |  | datasets'' where there is no measured activity, as highlighted in the |
|  |  | data fusion use case. Inclusion of such datasets can be useful as |
|  |  | they provide an opportunity to directly highlight untested regions of |
|  |  | chemical space. When multiple datasets are included in the data fusion, |
|  |  | some with more accurately measured activities, it provides the opportunity to increase |
|  |  | confidence in noisy data by merging data for scaffolds across the |
|  |  | datasets. |
|  |  |  |
|  |  | Compared to partitioning-based clustering methods (e.g., complete |
|  |  | linkage clustering), the overlapping scaffold approach described here |
|  |  | avoids the phenomenon of similar molecules being arbitrarily assigned to |
|  |  | exclusive clusters. Instead, using overlapping scaffolds, will |
|  |  | consider such molecules in the same group. |
|  |  |  |
|  |  | Given the different ways to generate scaffolds and to compute |
|  |  | overlapping scaffolds, a quantitative approach to characterizing |
|  |  | differences in these approaches is necessary. The use of common |
|  |  | proportion ($\textrm{CP}$), fragment efficiency ($\textrm{FragEff}$) |
|  |  | and proportion of information unique to a method ($\textrm{PIU}$) |
|  |  | places such differences within a sound statistical framework, allowing |
|  |  | for an objective comparison of fragmentation methods for a given |
|  |  | screen. They also extend the applicability of |
|  |  | methods to compare the output of different partitioning clustering |
|  |  | methods such as **\citeauthor**{Torres2009}, allowing them to be used for |
|  |  | non-overlapping fuzzy clusters. |
|  |  |  |
|  |  | For large structure collections, the |
|  |  | primary bottleneck is the fragmentation step. However, scaffold |
|  |  | generation is usually a one-time procedure. Furthermore, after |
|  |  | scaffold generation, the compound-scaffold associations reduce to |
|  |  | fixed time lookups and thus assesing scaffold membership (**\eg** |
|  |  | computing structure groups or activity groups) is very fast. |
|  |  |  |
|  |  |  |
|  |  | We hope our use of anecdotal and statistical methods to quantify |
|  |  | similarities and differences has convinced the reader that there is no |
|  |  | one fuzzy clustering method that is superior to others in all |
|  |  | situations. Instead, different methods bring nuances and slightly |
|  |  | different sets of SAR insights to the table, and the reader may |
|  |  | evaluate which performs the best on their own dataset. At the same it |
|  |  | is important to realize the importance of a chemists intuition and |
|  |  | experience, which can often bias scaffold and subsequent lead |
|  |  | selection. Thus a chemist with a ``favorite'' scaffold may |
|  |  | (unconciously) tend to downplay or ignore other scaffolds. While this |
|  |  | is unavoidable, visual and data-driven approaches such as described |
|  |  | here may help to combat such biases. |
|  |  |  |
|  |  | Our method for quantifying structure-activity overlap via NormCP is |
|  |  | novel, though similar in spirit and purpose to local hit rate |
|  |  | non-overlapping fuzzy clusters. Furthermore quantifying |
|  |  | structure-activity overlap using $\textrm{NormCP}$ is a novel |
|  |  | contribution, though similar in spirit and purpose to local hit rate |
|  |  | calculations that have been proposed for HTS triage**\cite**{Posner2009}, |
|  |  | with a comprehensive structural neighbor metric based on overlapping |
|  |  | scaffolds. We hope it will be explored more as a means of HTS triage and hit |
|  |  | set prioritization. |
|  |  |  |
|  |  | scaffolds. |
|  |  |  |
|  |  | In summary, the combination of anecdotal and statistical methods to |
|  |  | compare scaffold schemes and the resultant analysis of HTS datasets |
|  |  | highlights the fact that no single fuzzy clustering method is optimal, |
|  |  | and the most appropriate approach should be selected based on the |
|  |  | types of analyses described here. It is received wisdom among screening scientists |
|  |  | that ``chemical intuition'' plays a role when examining scaffolds, yet such intuition |
|  |  | can lead to biased selections of scaffolds and subsequently of |
|  |  | leads. The data driven approach coupled with intuitive visualizations, |
|  |  | as described in this study, provides a way to combat such biases. |
|  |  |  |
|  |  |  |
|  |  | \begin{acknowledgement} |
| ... | ... |  |

[**spotfire/chemblntd\_gsk\_TCAMS\_Stats\_PropByCompd\_eLNB.dxp**](https://spotlite.nih.gov/ncats/scaffoldanalytics/compare/e2325c81...master#b43966ef3e4964e9ba868d862b57c26710e40b19) 0 → 100644

[View file @7fca65b](https://spotlite.nih.gov/ncats/scaffoldanalytics/blob/7fca65b4814833372e08489c48859581ecb70925/spotfire/chemblntd_gsk_TCAMS_Stats_PropByCompd_eLNB.dxp)

This file is too large to display.