Cloud hosted APIs for cheminformatics designed for real time user interfaces

Alex M. Clark, Ph.D.

March 2014

© 2014 Molecular Materials Informatics, Inc.

http://molmatinf.com

Data Regimes

- Differences in kind based on size:
  - **small**: <1000 molecules; document-sized
  - **medium**: <100K; filesystem, heavy duty
  - **large**: database servers; limited operations
- Nimble client (mobile apps, web) either:
  - operate on **small** collections
  - limited window onto **large** collections
- Mobile+cloud workflows with **medium** size…

Overview

- Describing a workflow for tuberculosis; doing scaffold analysis, model building

- Split into:
  - **mobile apps** as the user interface
  - **cloud-hosted** algorithms for hard work
  - **desktop-based** sections for medium data

- Mobile+cloud very convenient for small data, and for well established tasks

- Desktop still primary for method development

TB Mobile

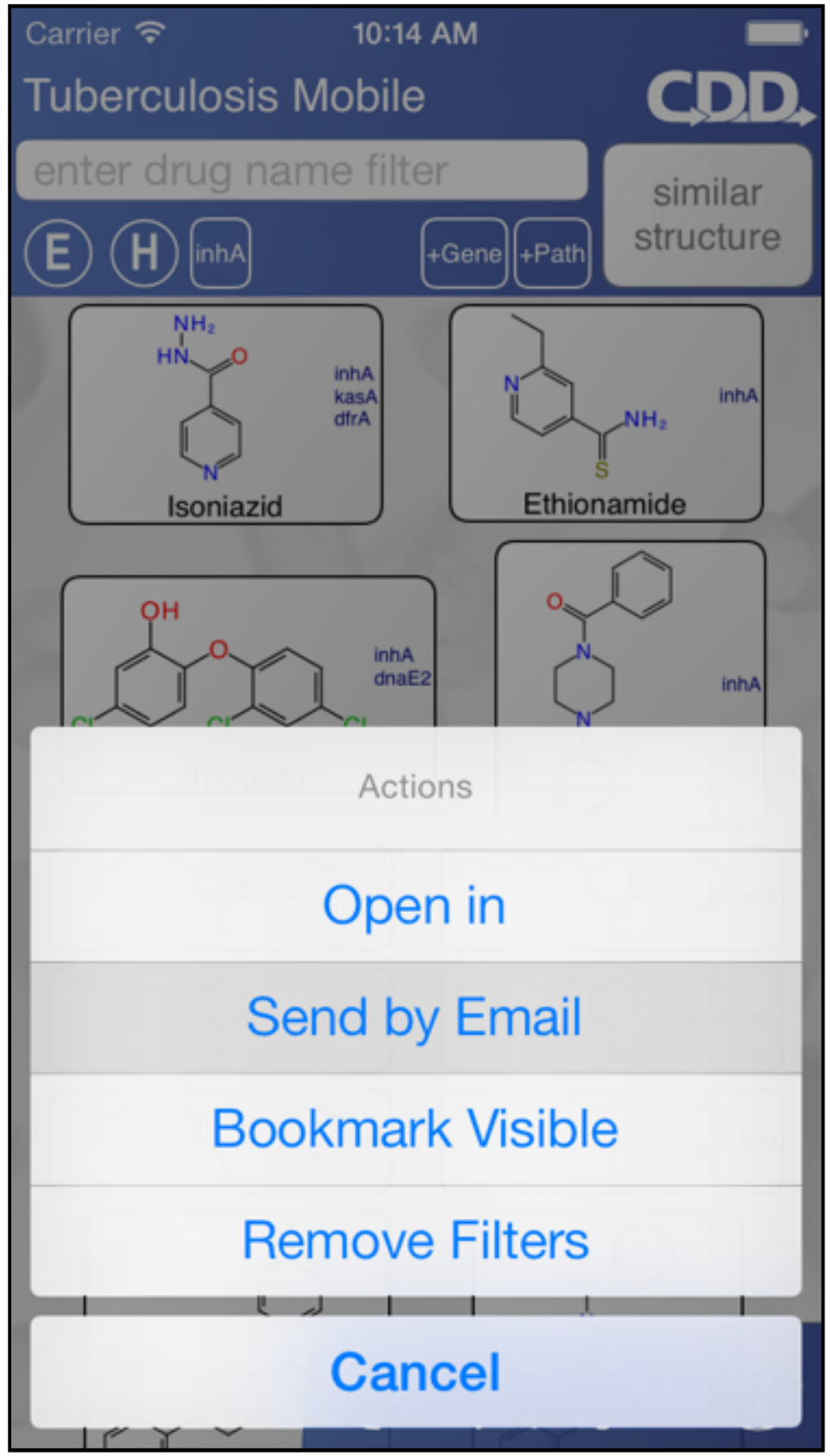
- Begins with a mobile app:

  - ~90 curated targets
  - ~ 800 molecules

- TB inhibition data abundant, but mostly no target info

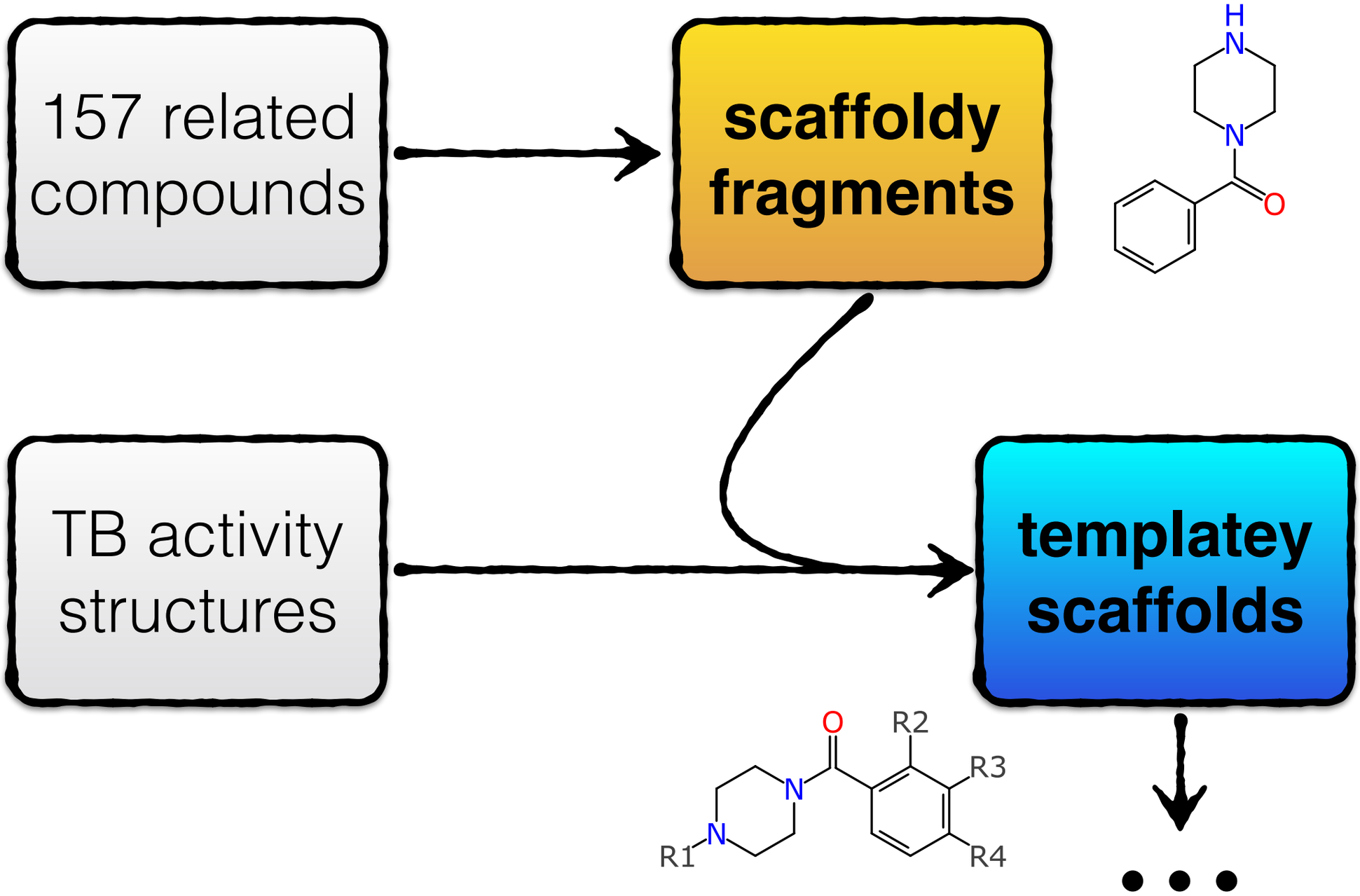
- Want all the actives against the **inhA** target (157)

- Generate leads using scaffold analysis

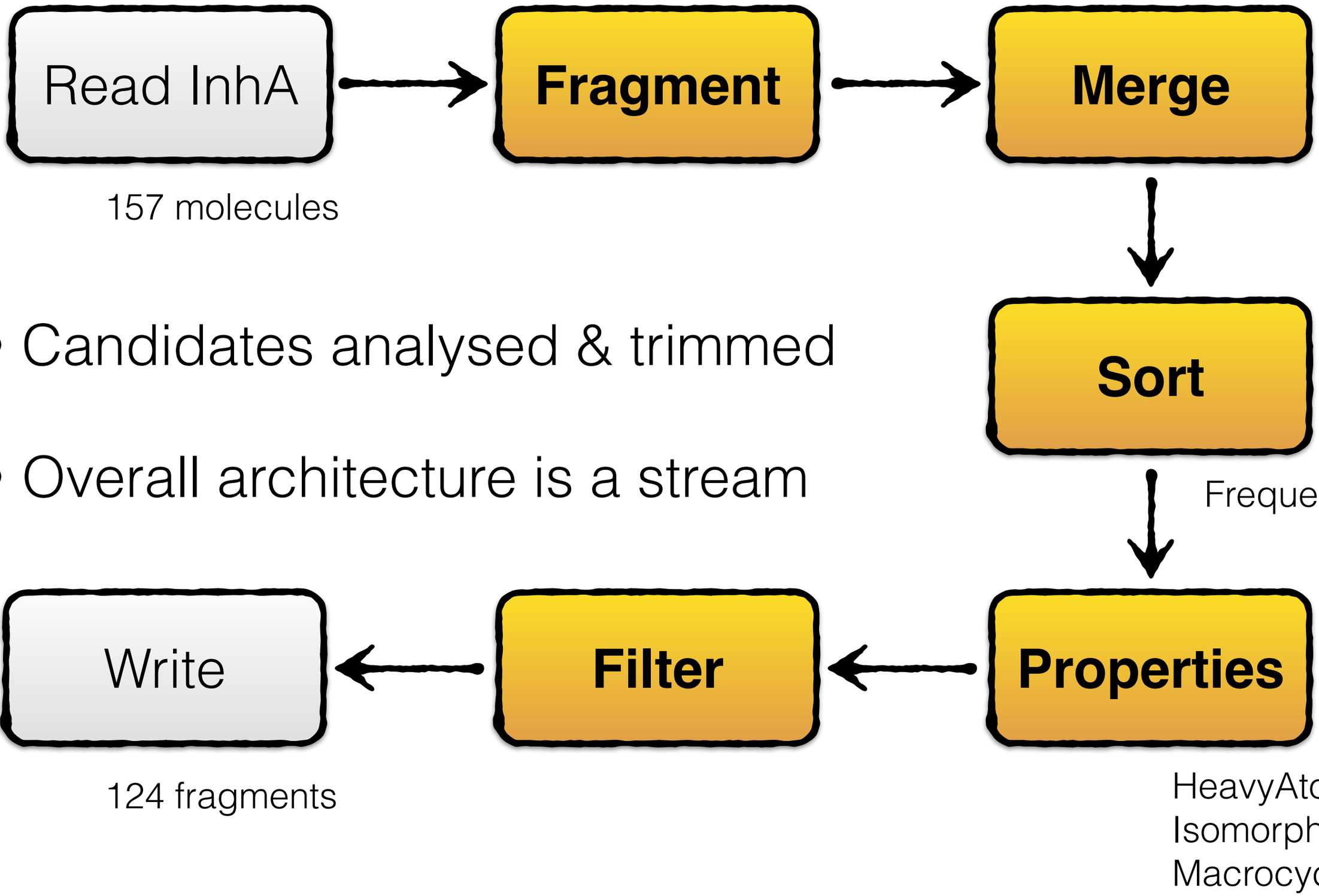


Scaffold Fragments

- What medicinally relevant scaffolds to use?



Filtering Scaffold Candidates



- Candidates analysed & trimmed
- Overall architecture is a stream

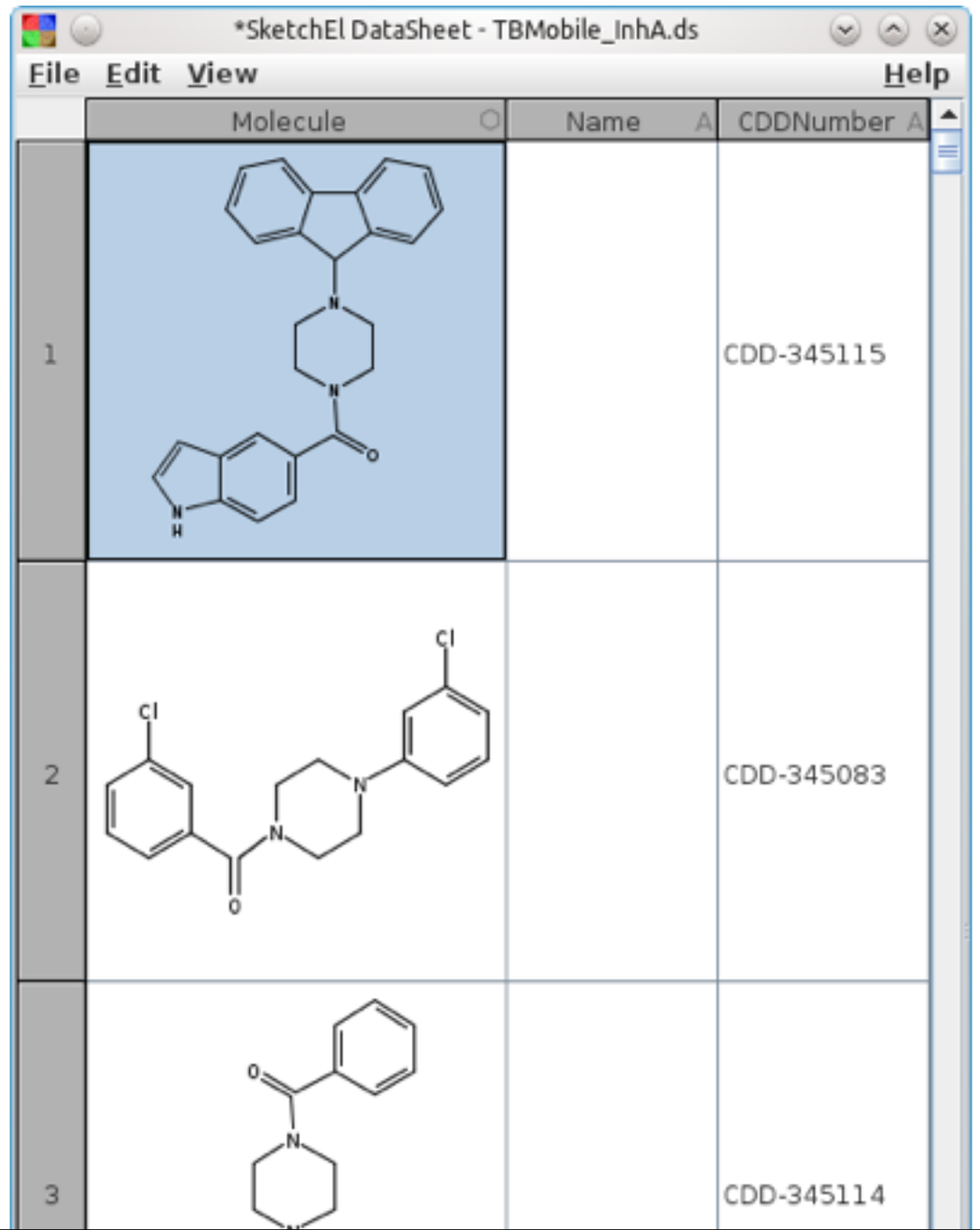
Pipelining

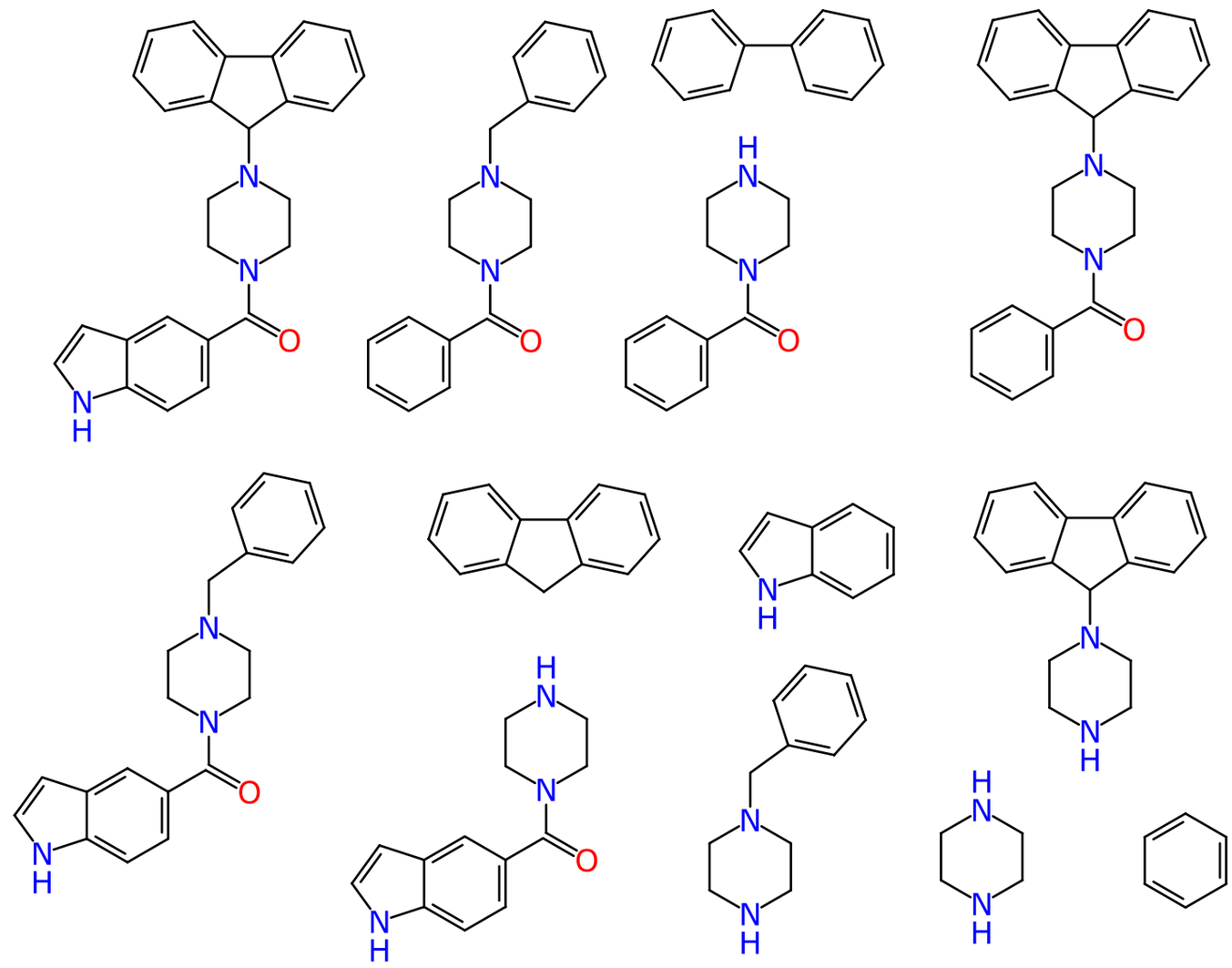
- Not quite cloud (yet)
- Infrastructure for streaming nodes together: build workflows using a script
- Roadmap: build selected workflows, out of prepackaged nodes
- Expose as webservices: for use by mobile apps


```
{
 "op":"com.mmi.core.op.CollapseUnique",
 "id":102,
 "name":"Collapse",
 "parameters":
 {
  "keyColumn":"Molecule",
  "countColumn":"Degeneracy",
  "collapseColumn":["Target"],
  "collapseOperator":[","]
 },
 "inputs":[[101,1]],
 "outputs":1
},
{
 "op":"com.mmi.core.op.Sort",
 "id":103,
 "name":"Collapse",
 "parameters":
 {
  "columns":["Degeneracy"],
  "directions":[-1]
 },
 "inputs":[[102,1]],
 "outputs":1
},
{
 "op":"com.mmi.core.op.MoleculeProperties",
 "id":104,
 "name":"Properties",
 "parameters":
 {
  "heavyAtoms":"HeavyAtoms",
  "isomorphisms":"Isomorphisms",
  "macrocycles":"Macrocycles"
 },
 "inputs":[[103,1]],
 "outputs":1
},
{
 "op":"com.mmi.core.op.FilterProperties",
 "id":105,
 "name":"Filter",
 "parameters":
 {
  "name":["HeavyAtoms","Isomorphisms","Macrocycles"],
  "operator":[">=","<=","="],
  "value":[10,4,0]
 },
 "inputs":[[104,1]],
 "outputs":1
},
```

Fragmentation

- Consider each structure: break it into pieces, enumerate scaffold-like fragments



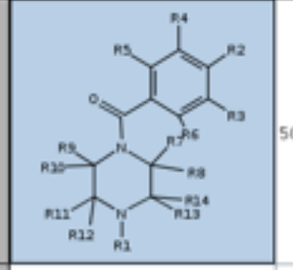
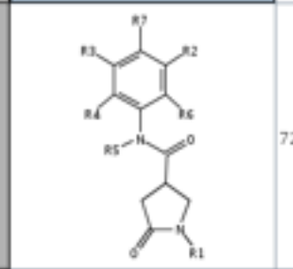
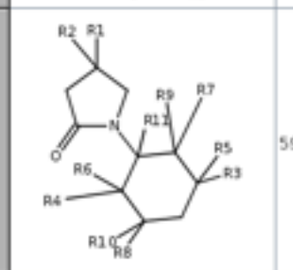
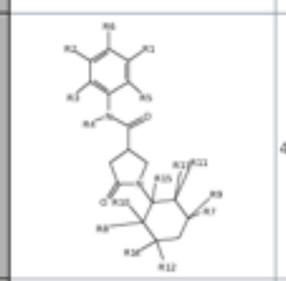
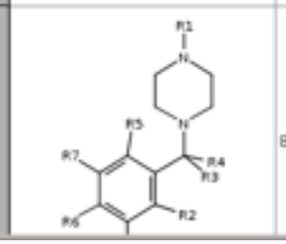


Decorating

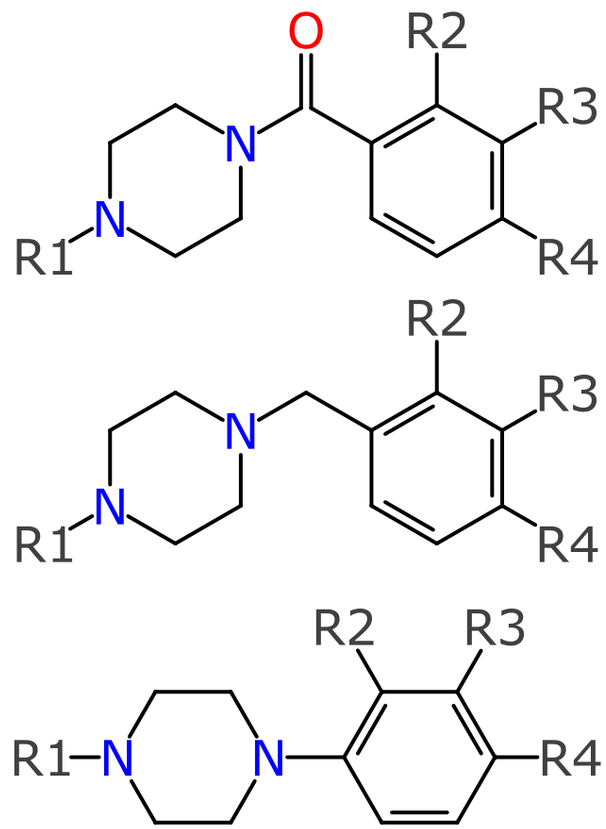
- Have scaffoldy fragments, 5425 measurements

*SketchEl DataSheet - scaffold_templates.ds

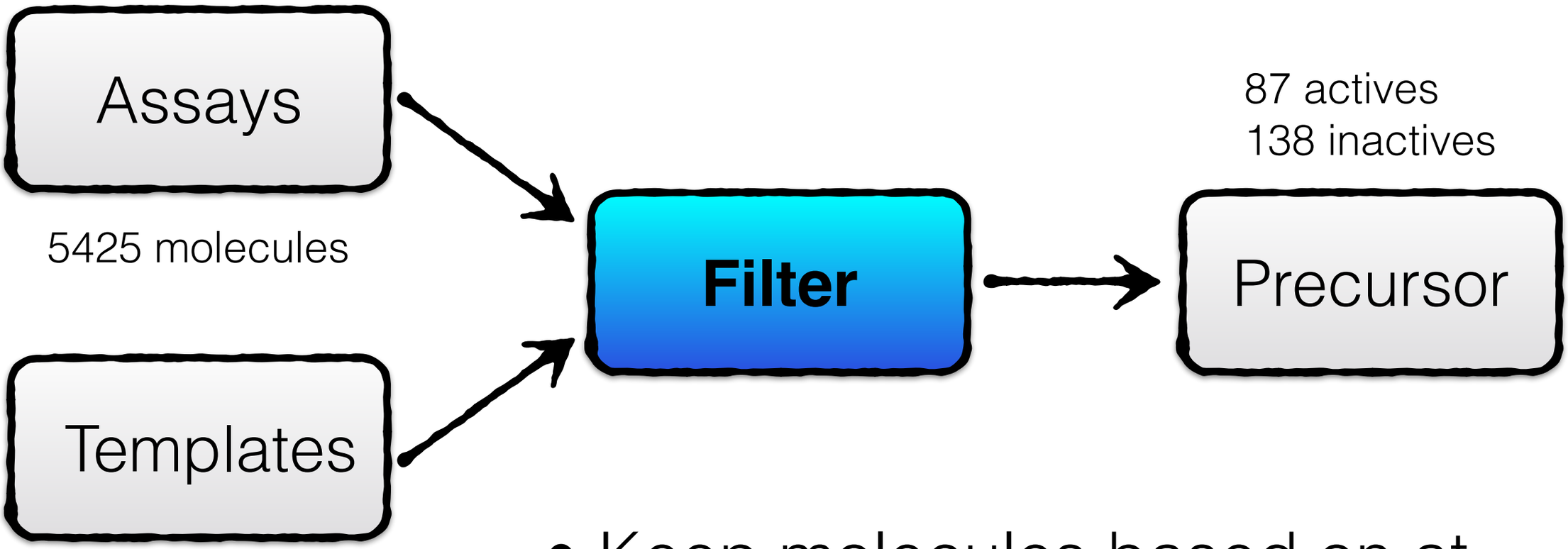
File Edit View Help

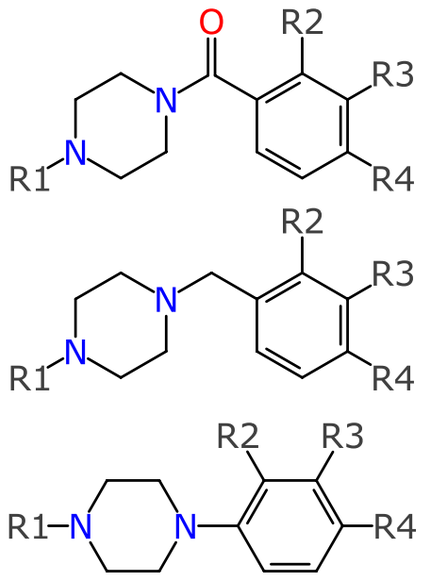
	Scaffold	Matches	Degeneracy	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15
1		50	536	536	344	48	48	10	10	4	4	4	4	16	16	16	16	0
2		72	144	144	70	70	18	2	18	48	0	0	0	0	0	0	0	0
3		59	2172	1086	1086	192	48	192	48	48	192	48	192	384	0	0	0	0
4		49	1464	778	778	56	4	56	612	96	32	96	32	32	96	32	96	192
5		83	924	916	124	264	264	124	360	96	96	0	0	0	0	0	0	0

- Do a trial matching: templates & stats



Scaffold Selection





- Keep molecules based on at least one template
- Output is suitable for the next stage in the workflow

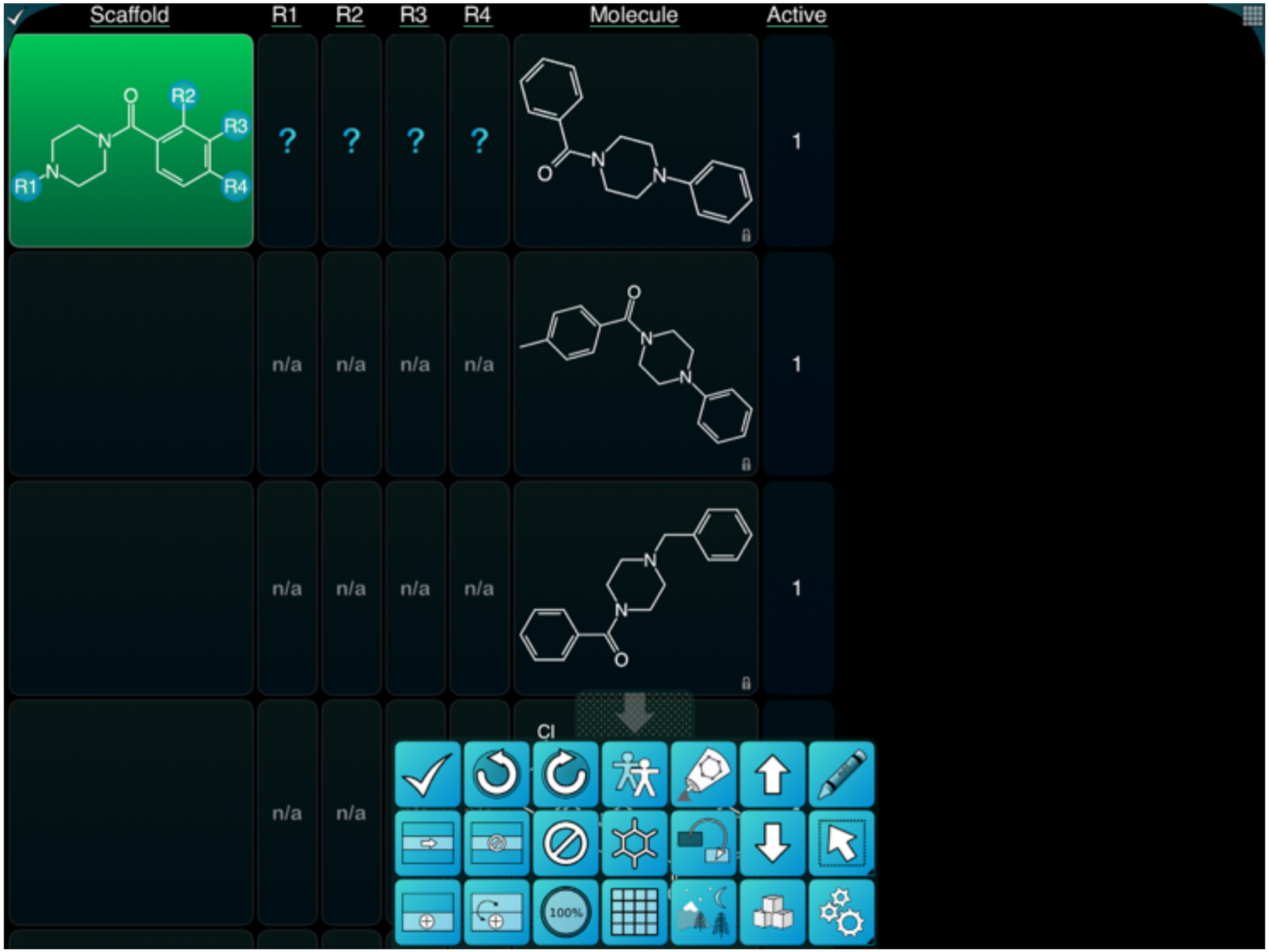
SAR Table App

- Back to mobile apps: want to deliver the 225 compounds to iPad/iPhone…

  - email
  - dropbox
  - web

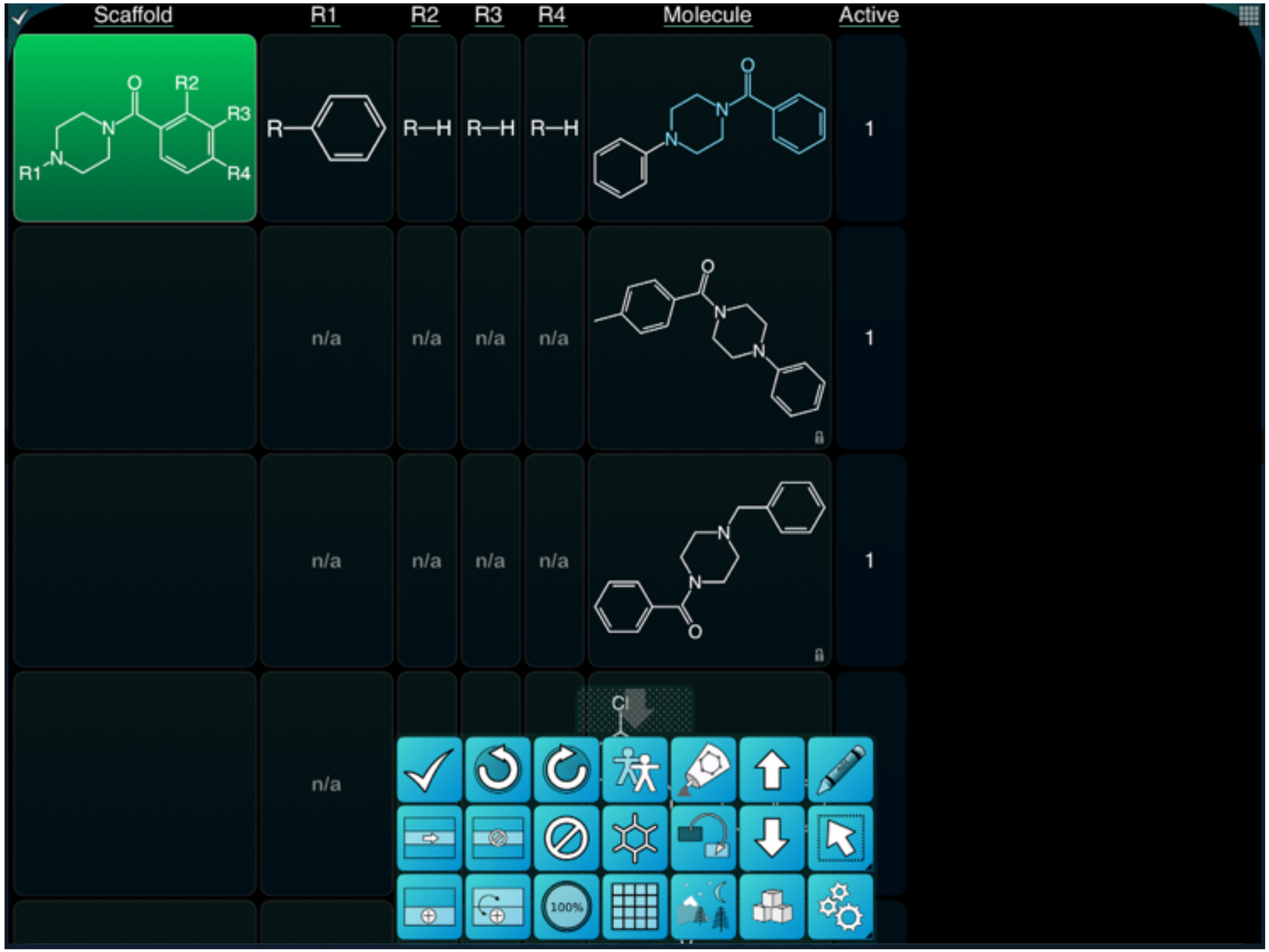
- **SAR Table** app designed for small documents: content creation, focused analysis, and cloud-assisted functions

Import



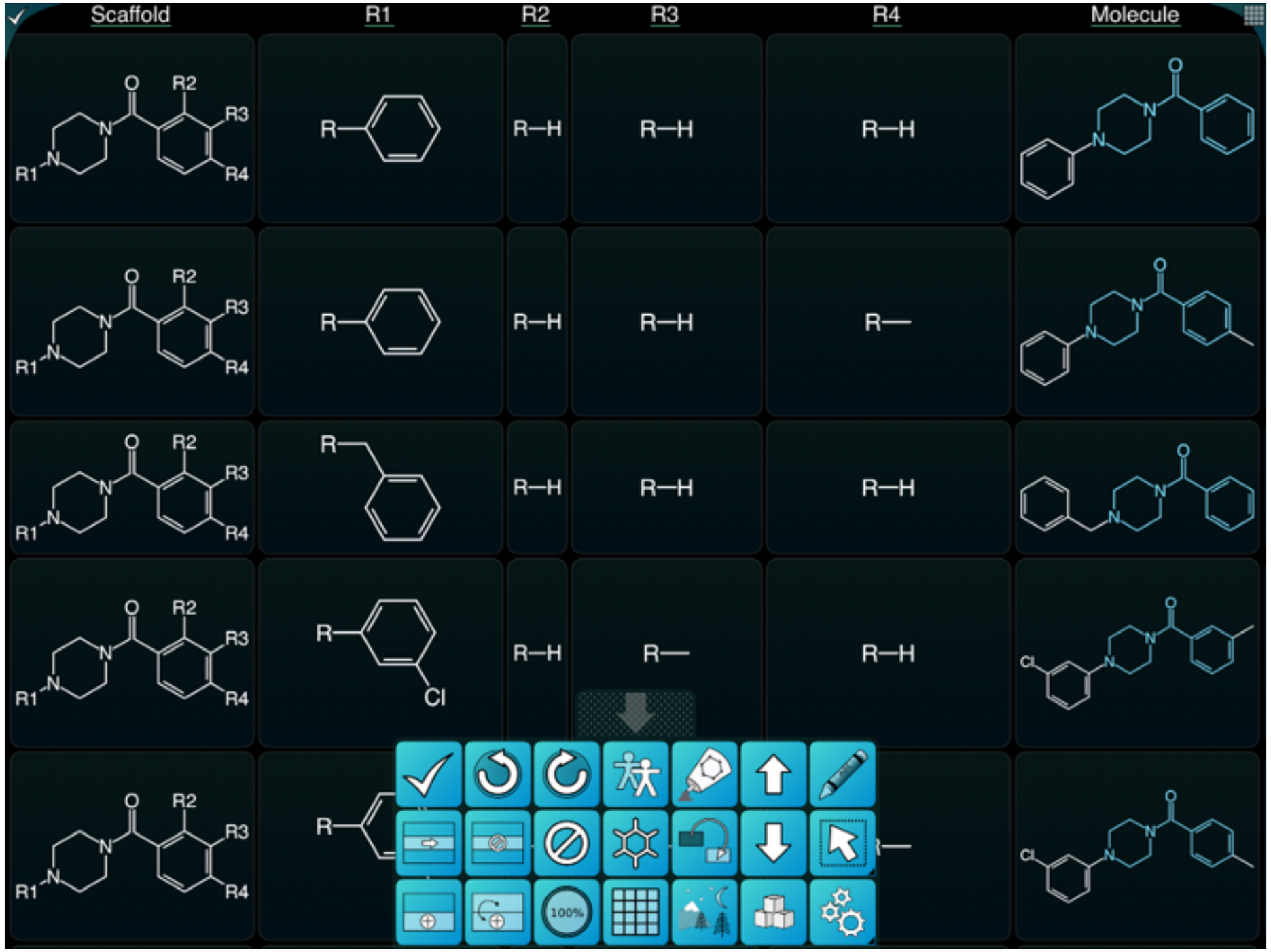
- Launch datasheet, draw first scaffold…

Scaffold Assignment



- Ask the webservice to assist: complex, fast

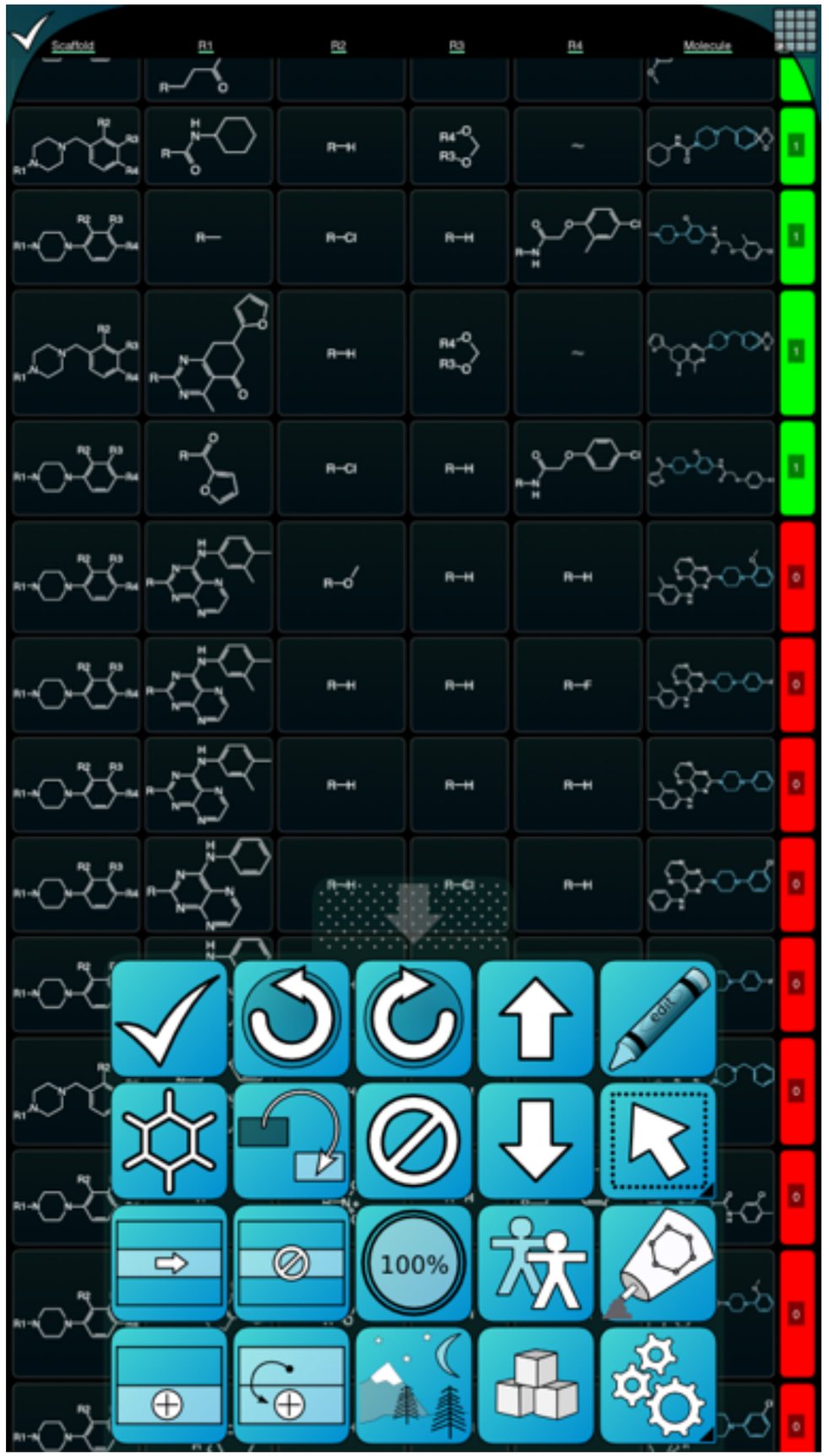
Multi-Scaffold Assignment



- Assign scaffolds in bulk: complex, quite fast

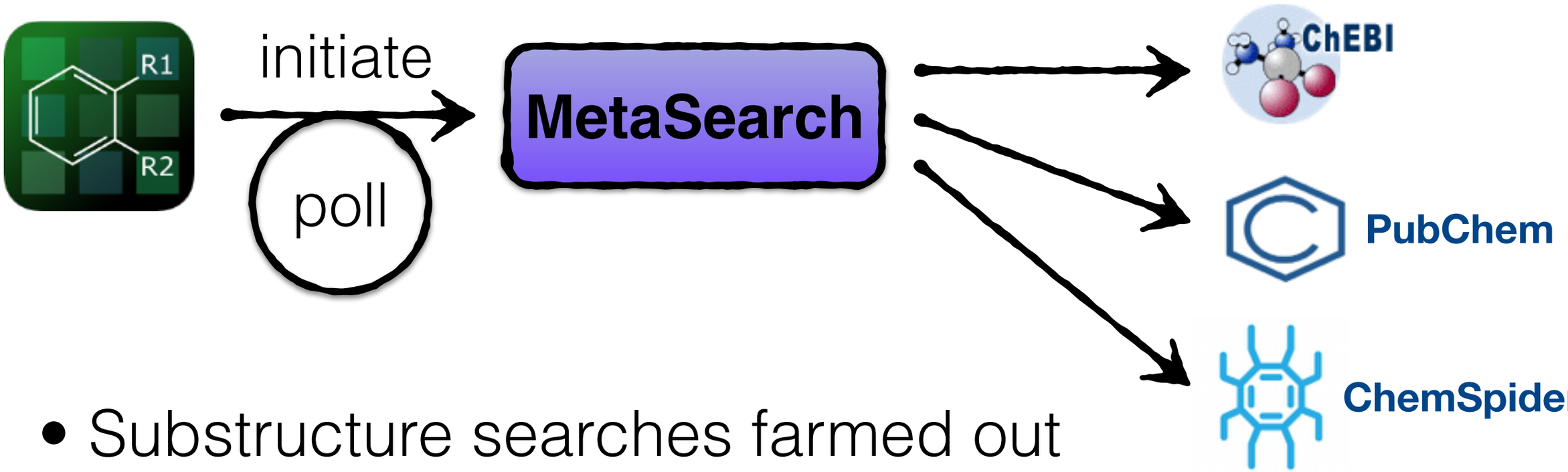
More Data

- Have scaffolds and substituents assigned
- Can gain valuable insight just from that
- What about **public databases**: what else do our 3 scaffolds match?



Searching

- Search for a template; optionally narrow substituent values; want only new compounds



- Substructure searches farmed out to well known **large** data services

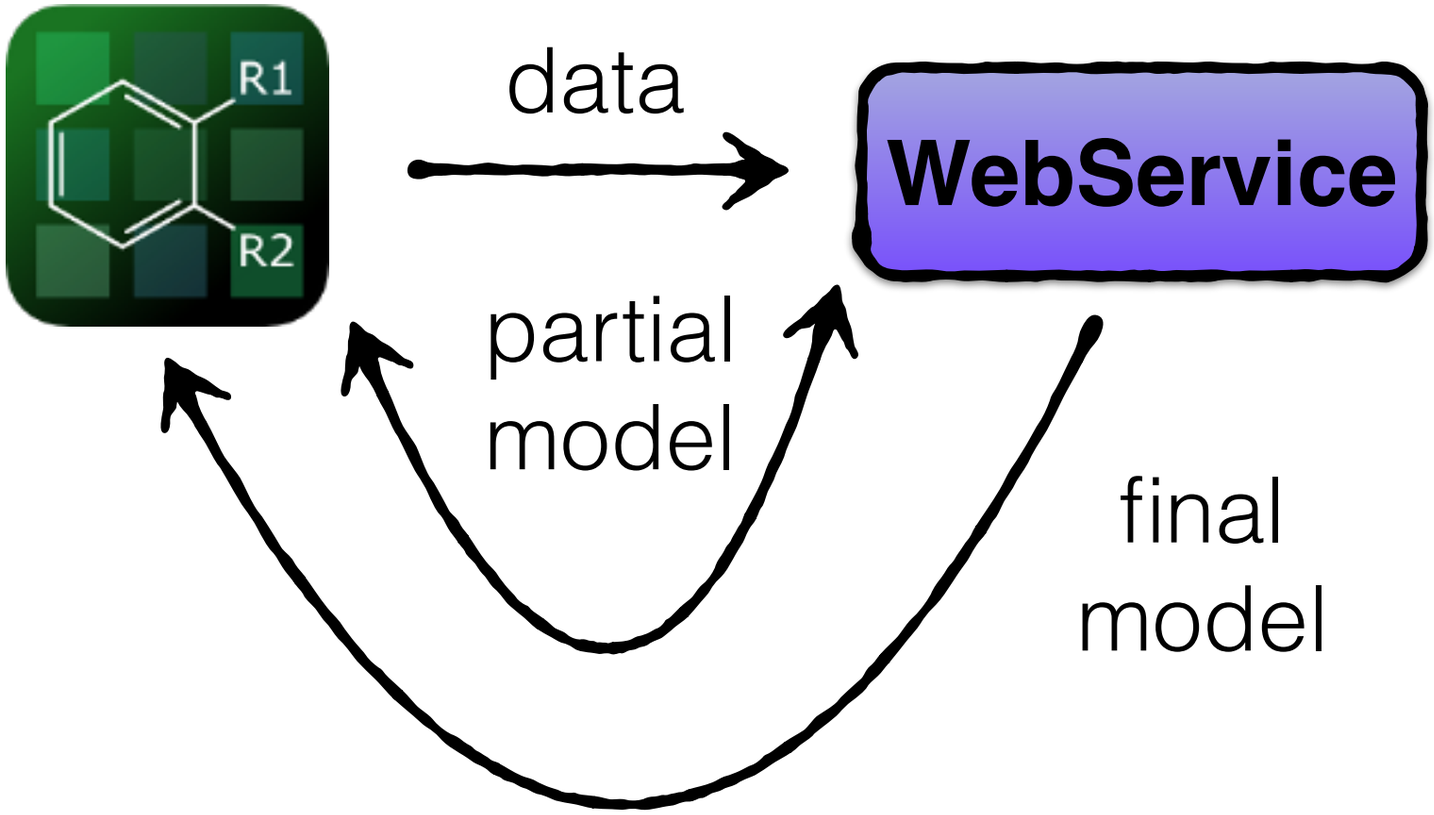
- Middleware post-processes with scaffold analysis & assignment

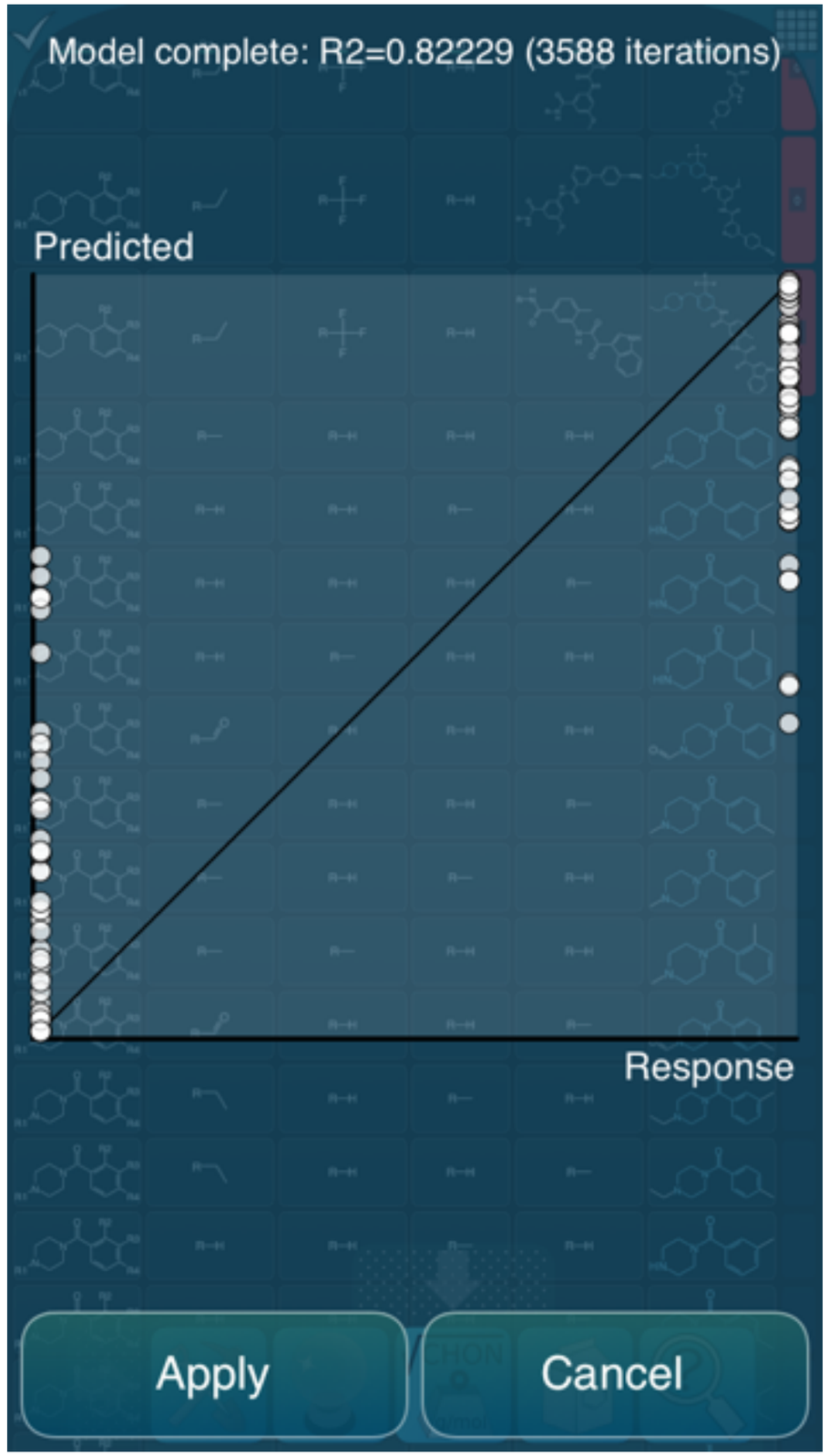
Results

- Results are marked up
- Uses existing fragments for context
- No duplicate structures
- All compounds are known…
- … can be made or purchased.

Model Building

- Use structures with known activities to create a structure-activity model

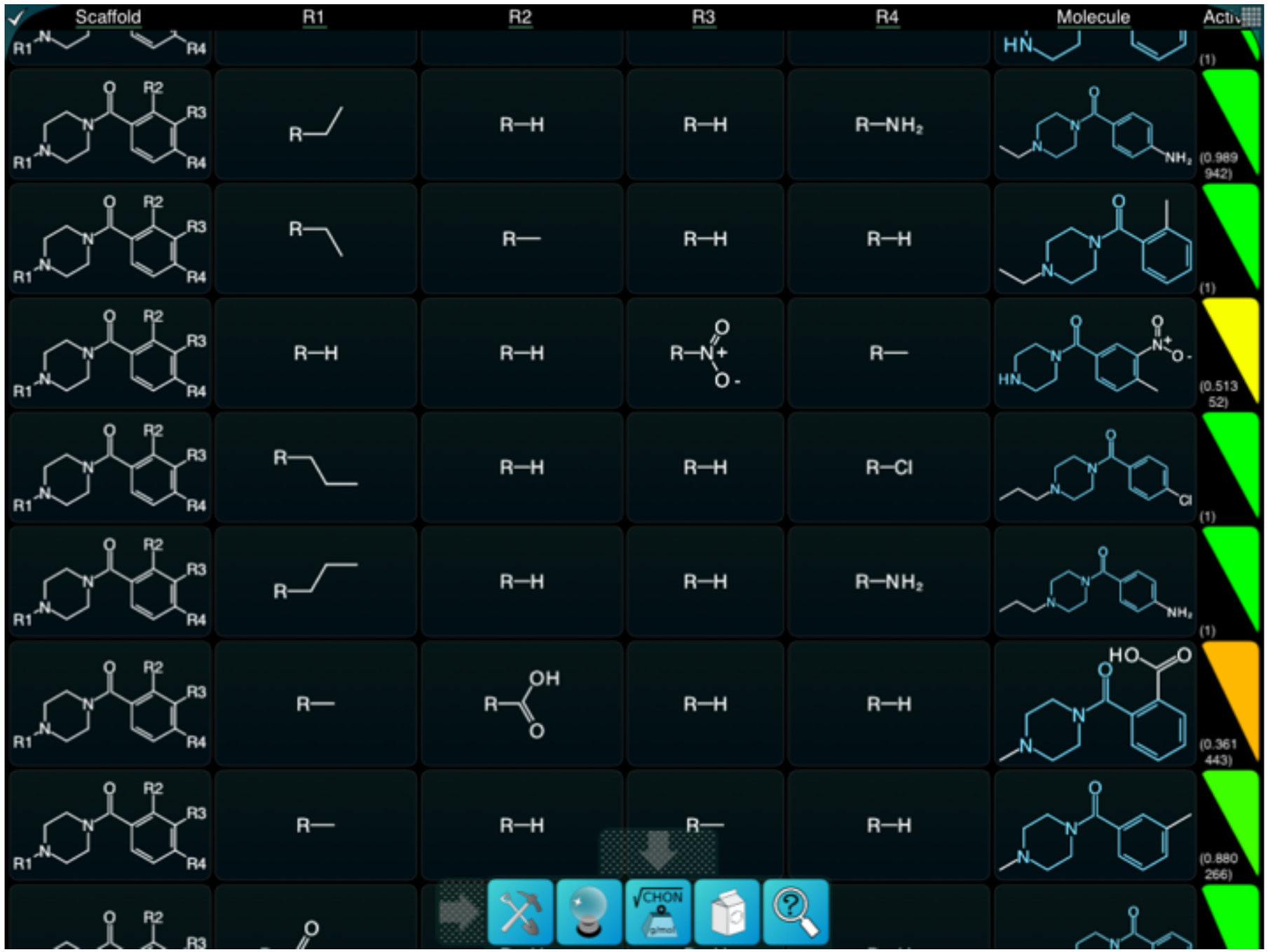




- **Slow** calculation, **small** data

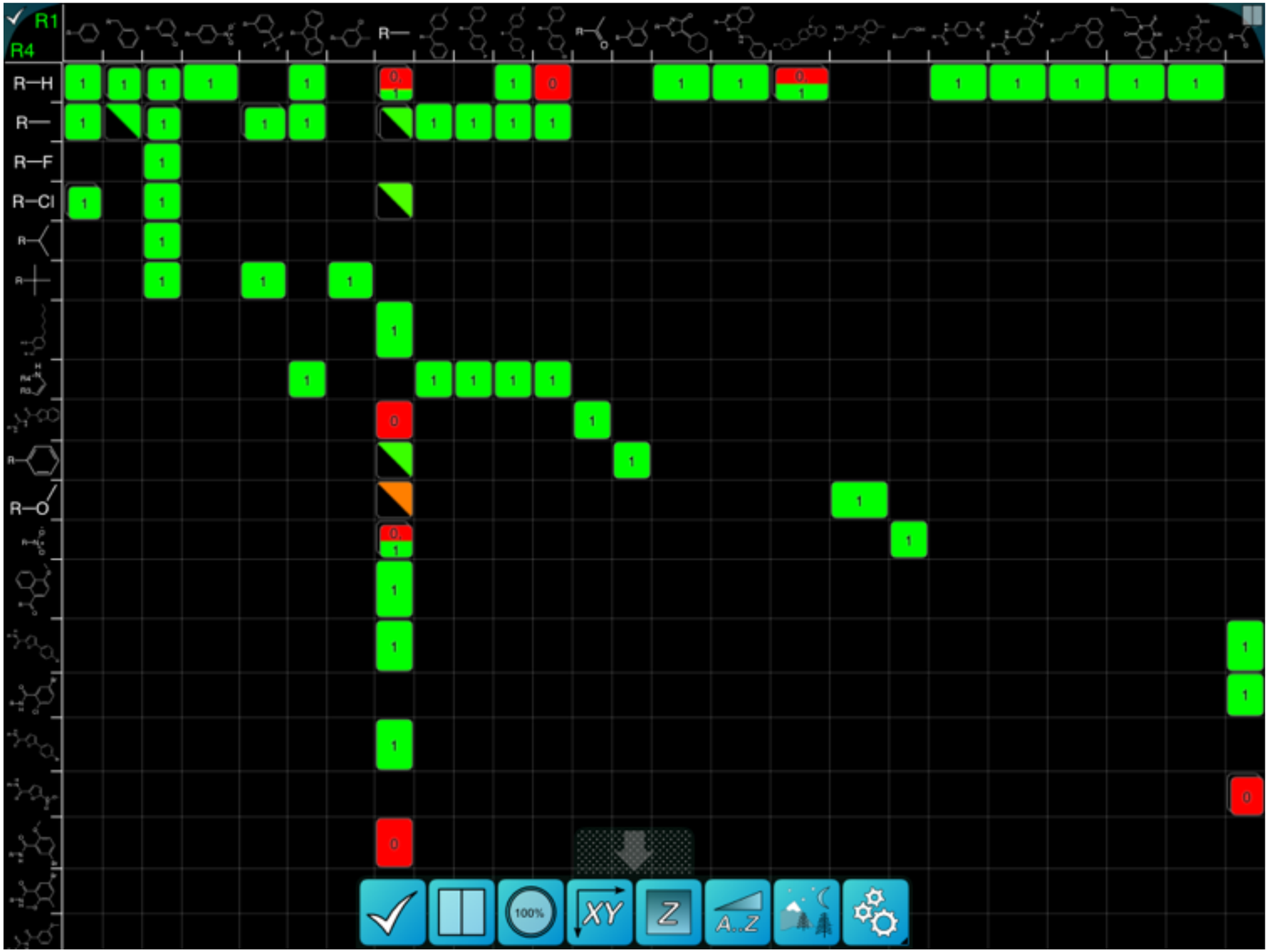
Model Application

- Predicted activities for looked-up compounds…



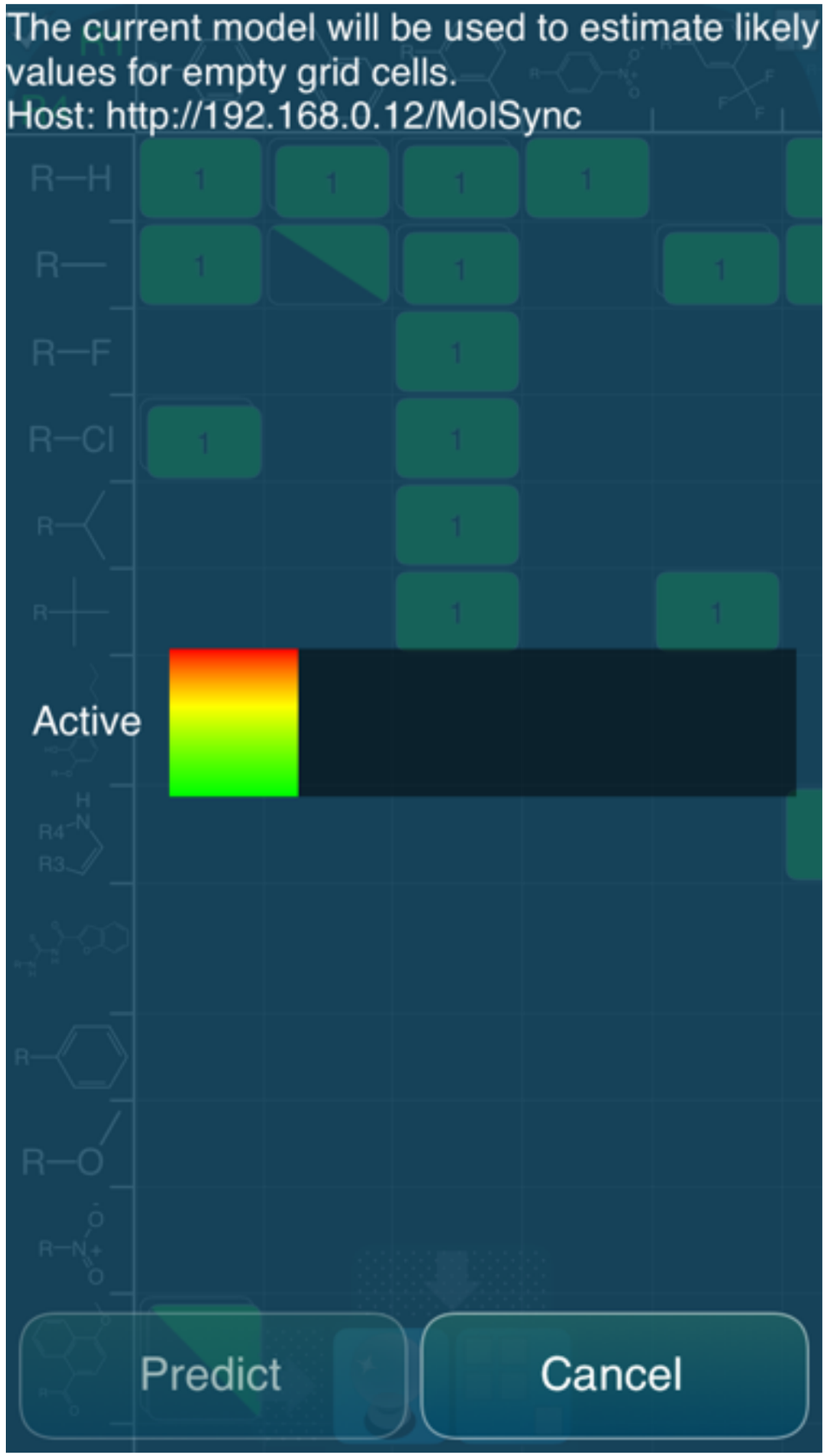
Matrix View

- Plot **R1** vs **R4**: examine second order SAR



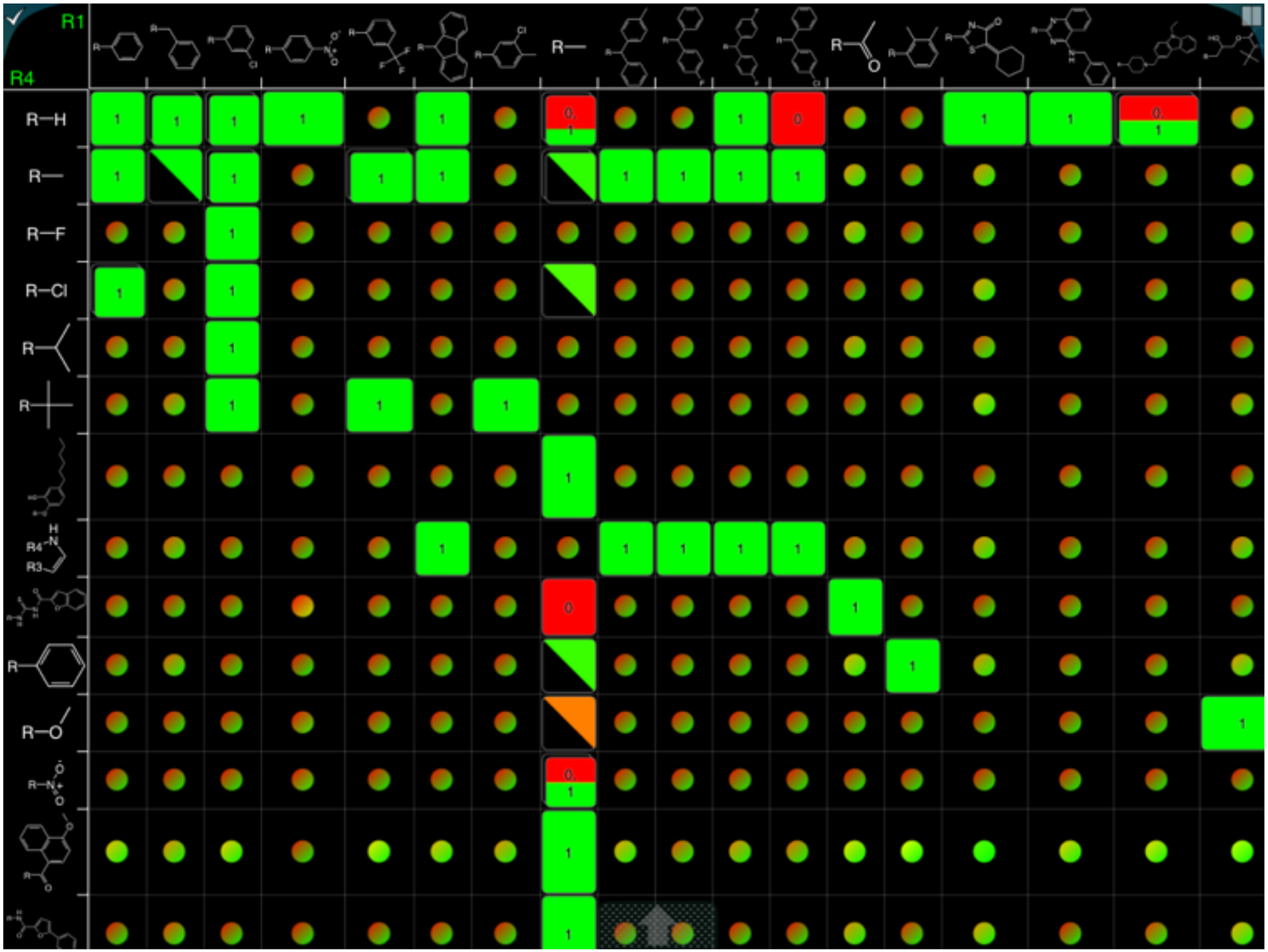
Filling in Blanks

- Each blank cell: create & score chimeric structures
- Gather distribution of activities
- Total calculation: slow
- Performance: overhead amortised in blocks (e.g. 10 cells per request)



Matrix Predictions

- Shows **measured**, **available** & **hypothetical**…



Conclusion

- Mobile + cloud can accomplish many sophisticated tasks
- Stateless webservices very easy to deploy
- Work on **small** datasets, use **large** databases
- **Medium** sized data is problematic
- Can fallback to desktop: facile communication
- Apps & webservices very well suited to mature workflow tasks

Acknowledgments

- Sean Ekins, Barry Bunin & CDD
- RSC & ChemSpider, PubChem, ChEBI
- Inquiries to **info@molmatinf.com**

MOLECULAR MATERIALS INFORMATICS

http://molmatinf.com
http://molsync.com
http://cheminf20.org

@aclarkxyz