Cheminformatics from the end-user perspective: Past, present and future.

Paul Greenspan,
Senior director of oncology chemistry, Takeda

April 11, 2016
Thank you in advance for your patience!

• This will NOT be a technical presentation. Sorry about that!

• I have been an industrial medicinal chemist for 25 years.
  – 11 years at Novartis (arthritis, inflammation)
  – 14 years at Millennium/Takeda (oncology)

• I am definitely NOT a cheminformaticist or computational chemist, but I have a lot of interest in the field, and greatly appreciate the value.

• Today, I will present my perspective on the evolution of cheminformatics over the course of my career, and what key challenges lie ahead.
What were things like 25 years ago for a medicinal chemist?

- Typical chemistry throughput might be 10 compounds/chemist/month
- An “HTS” might be 10,000 compounds/month

- Very limited use of assays beyond primary screens.
  - 1 or 2 datapoints per compound.

- What was the state of “cheminformatics” 25 years ago?
  - Medicinal chemistry databases were just being introduced
  - MDL was the only game in town
  - Most project teams kept assay data in private databases (or spreadsheets)
  - Until ~2000, the key challenge was getting data into a searchable database
Remember when this was state of the art?

- Customizable GUI, multiple display options, structure and data searching
Volume of data has exploded in past 25 years

**More compounds**
- Routine HTS screens of $>10^6$ compounds
- High-throughput synthetic chemistry
- New ultra-high-throughput screening approaches (e.g. DNA-encoded libraries)
- Enormous “virtual” compound libraries.
- External vendors with vast catalogs of compounds

**More data per compound**
- Extensive cross target selectivity screening
- Broad target-class screens (e.g. Kinome panels)
- Routine HT predictive ADMET screening
- Predictive modeling generating lots of “virtual” data
- Large external chem/biology databases (pubchem, chembl, etc.)
Cheminformatics has come a long way…

• Global, user-friendly chemistry/biology databases are commonplace (if not universal)

• Predictive modeling has become much more mainstream

• Broad implementation of electronic notebooks has made even “raw data” accessible.

• Entirely new ways of analyzing data have taken hold:
  – Dynamic querying and visualization tools (spotfire, etc.)
  – Multi-parameter optimization methodologies allow more “holistic” analysis
  – Specialty tools (MMP, activity landscape analysis, etc.)
  – Clustering, framework analysis
Chemistry Dashboards integrate data seamlessly

Example: Dotmatics Vortex
These changes have redefined the challenge in fundamental ways

• 25 years ago, the goal was to make data available to allow chemists to review SAR data manually.
  – We couldn’t envision tools to allow for more than that.
  – The datasets were small and simple enough to make this practical

• Today, datasets are far too large and complex for chemists to consume, analyze and draw conclusions manually from the data they receive.

• The key cheminformatics challenge is to enable chemists to make optimal use of all this data:
  – Construct testable hypotheses
  – Effectively prioritize design ideas
  – Assist chemists’ imagination in generating new approaches
The Great Computational/Med Chem Divide

There are several challenges in supporting med chemists in working with large datasets:

• Med chemists don’t like math!
  – We tend to think visually, rather than mathematically.
  – Outcomes of statistical analyses must be conceptually straightforward.

• Chemists don’t deal well with uncertainty:
  – A chemical structure is absolute. Biological data is not.

• There is no perfect way to parameterize a chemical structure:
  – Chemists may not agree with calculated similarities, clustering, etc.
  – Meaning of atom connectivities can be very context-dependent.
Dumbing down the data

- If chemists don’t like math, and struggle to conceptualize large datasets, then let’s keep it simple.

- Create “rules” that any idiot can obey:
  - Lipinski Rule of 5.
  - Internal cut-offs imposed by many pharma organizations

- But can this possibly be right?
  - Aren’t these things context dependent?
  - Is MW of 495 really infinitely better than MW of 505?
  - If lipophilicity is low, couldn’t we back off on our MW cut-off?
Is there a better approach?

- Unintuitive mathematical constructs have limited appeal.
- Oversimplification can lead to erroneous decision-making
- Datasets are too large and complex to expect a chemist to retrieve all potential value through manual inspection.

- How do we help chemists in a way that plays to their strengths?
  - Data visualization
  - Computational identification of data “gems”
Visualization: a big breakthrough

- Spotfire introduced the concept of interactive visualization to medicinal chemistry and drug discovery
  - Bridged the gap between manual SAR analysis and statistical methods.
  - Allowed chemists to be in control: view data from variety of perspectives, pose questions that can only be answered with aggregate data.
  - Outputs are visual, not mathematical.
  - Allowed for real-time, iterative data interrogation and hypothesis generation
Example: No obvious trends across data-set
Is there a trend if we only look at amines?

- chemistry queries with visual output
How about amines with logp < 3?

- Explore additional data relationships interactively
- Create testable hypotheses
Another Breakthrough: finding the data “gems”

- Sometimes, the most important data is “small”:
  - The comparison of a few datapoints may tell a critical story

- But how do chemists pick that out from all the noise?

- Cheminformatics has helped chemists to home in on key data:
  - Matched molecular pair analysis
  - Activity landscapes
### The Power of Matched Molecular Pair Analysis

**Dossetter, et. al.** *Drug Discovery Today*, Vol. 18, p. 724

#### Table 1: Assay Endpoints

<table>
<thead>
<tr>
<th>Assay Endpoint</th>
<th>n</th>
<th>Δ Mean</th>
<th>SE</th>
<th>SD</th>
<th>% Improve</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>log $D_{50}$ (all H to F)</td>
<td>9902</td>
<td>-0.10</td>
<td>&lt;0.01</td>
<td>0.34</td>
<td>24 (1.6 fold)</td>
<td>Papadatos [20]</td>
</tr>
<tr>
<td>log Aq. Solubility (all H to F)</td>
<td>4273</td>
<td>-0.10</td>
<td>&lt;0.01</td>
<td>0.37</td>
<td>17 (2 fold)</td>
<td>Papadatos [20]</td>
</tr>
<tr>
<td>Cytochrome P450 inhib pIC$_{50}$</td>
<td>1A2</td>
<td>0.03</td>
<td>&lt;0.01</td>
<td>0.45</td>
<td>32 (1.6 fold)</td>
<td>Gleeson [19]</td>
</tr>
<tr>
<td></td>
<td>2C9</td>
<td>0.04</td>
<td>&lt;0.01</td>
<td>0.37</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2C19</td>
<td>0.04</td>
<td>&lt;0.01</td>
<td>0.37</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2D6</td>
<td>0.03</td>
<td>&lt;0.01</td>
<td>0.41</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3A4</td>
<td>0.07</td>
<td>&lt;0.01</td>
<td>0.42</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>Permeability (log nM/s)</td>
<td>2848</td>
<td>+0.01</td>
<td>&lt;0.01</td>
<td>0.32</td>
<td>15 (2 fold)</td>
<td>Gleeson [19]</td>
</tr>
<tr>
<td>Rat in-Vivo Unbound Clearance</td>
<td>96</td>
<td>0.11</td>
<td>0.04</td>
<td></td>
<td></td>
<td>Sutherland [21]</td>
</tr>
<tr>
<td>Potency at target classes pIC$_{50}$</td>
<td>Kinases (7)</td>
<td>3.8 (10 fold)</td>
<td>Hajduk [8]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(all H to F)</td>
<td>Class 1 GPCR (9)</td>
<td>3.4 (10 fold)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Others (14)</td>
<td>3.7 (10 fold)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>hERG pIC$_{50}$</td>
<td>1572</td>
<td>-0.10</td>
<td>0.01</td>
<td>0.34</td>
<td>24 (2 fold)</td>
</tr>
<tr>
<td></td>
<td>hERG pIC$_{50}$ (all H to F)</td>
<td>4243</td>
<td>18 (2 fold)</td>
<td>Papadatos [20]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Table 2: Assay Endpoints

<table>
<thead>
<tr>
<th>Assay Endpoint</th>
<th>n</th>
<th>Δ Mean</th>
<th>SE</th>
<th>SD</th>
<th>% Improve</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>log $D_{50}$</td>
<td>252</td>
<td>+0.18</td>
<td>0.44</td>
<td>4.9 (3.2 fold)</td>
<td>Dossetter [15]</td>
<td></td>
</tr>
<tr>
<td>log Aq. Solubility (from solid)</td>
<td>711</td>
<td>-0.22</td>
<td>0.02</td>
<td>0.36</td>
<td>34 (all &gt;0)</td>
<td>Leach [12]</td>
</tr>
<tr>
<td>Human PPB K$_i$</td>
<td>171</td>
<td>+0.06</td>
<td>0.02</td>
<td>0.29</td>
<td>65 (all &gt;0)</td>
<td>Leach [12]</td>
</tr>
<tr>
<td>Rat PPB K$_i$</td>
<td>407</td>
<td>+0.15</td>
<td>0.01</td>
<td>0.29</td>
<td>77 (all &gt;0)</td>
<td>Leach [12]</td>
</tr>
<tr>
<td>A2HLM log Cl$_{int}$</td>
<td>497</td>
<td>-0.06</td>
<td>0.02</td>
<td>0.36</td>
<td>8.6 (3.2 fold)</td>
<td>Dossetter [15]</td>
</tr>
<tr>
<td>Pr HLM log Cl$_{int}$</td>
<td>491</td>
<td>0.00</td>
<td>0.02</td>
<td>0.38</td>
<td>18 (2 fold)</td>
<td>Lewis [14]</td>
</tr>
<tr>
<td>Rat Oral Bioavailability log AUC</td>
<td>551</td>
<td>+0.09</td>
<td>0.03</td>
<td>0.65</td>
<td>55 (all &gt;0)</td>
<td>Leach [12]</td>
</tr>
</tbody>
</table>

**Potency:**
- Kinases
- GPCRs
- Solubility
- Rat PPB
- Hum PPB
- HLM
- Rat F AUC
- hERG

Need for Enhanced rigor with MMPA?

- 4 pairs sufficient to identify significant differences with homogenous data.

- 10-20 pairs needed if data comes from different assays.

Kramer et. al. J. Med. Chem. 2014, 57, 3786
Activity Cliff Pathways

- Vasopressin VIa data from CHeMBL
- Analysis capture key SAR inflection points
- Pulling this data manually out of a large database would be difficult or impossible.

Paradigm shift: Multi-parameter optimization

• Historically, chemists have relied on filters for decision-making
  – Selection of compounds for secondary, tertiary screening
  – Choosing compounds to synthesize or purchase.

• Very simple to implement and conceptualize

• Serious drawbacks:
  – Greatly exaggerates small differences in parameter values
  – Overly rigid: filter values not impacted by other parameters
  – Order of filters can have unintended consequences:
    • Good compound can be lost early if it barely misses the first filter.

• MPO allows chemist to take all parameters into account simultaneously
Marriage of visualization and MPO: Golden Triangle

- Attempt develop more robust model for PK optimization
- Case is made primarily through visualization of multi-dimensional data

Probabilistic Scoring in Stardrop

- Stardrop allows chemist to control parameter weighting and selection
- Visualization allows chemist to readily see impact of each parameter
Predictive modeling: then and now

• Pitfalls of predictive modeling in the 90’s:
  – Focus on building “global” models that try to explain everything.
  – Use of “opaque” statistical methods (PLS, PCA)
  – Lack of clarity regarding limits in predictiveness

• Predictive modeling fell out of favor:
  – Frustration of chemists who didn’t understand models, and couldn’t determine their limitations.
  – Backlash from “overhype” (companies overselling modeling software)
  – No good way to incorporate into chemistry workflow

• We are now seeing a resurgence in predictive modelling:
  – Better understanding of limitations and appropriate uses.
  – Greater focus on local models.
  – Visualization tools allow chemists to interact with models, and understand drivers of predictions
What has this innovation given us?

• Chemists can now effectively interrogate large datasets, discover trends, and form hypotheses.

• Chemists can find the "data gems" that could easily be lost in the noise of large data-sets.

• Chemists can apply predictive modeling to real-world problems, and understand when and how it can be used.

• Chemists can be much more sophisticated in prioritization and decision-making
So, what are the next challenges?

• Better utilization of external data:
  – Integration of large external databases with internal tools.
  – Effective means of handling heterogeneous data-sets.
  – “Real-time” data extraction and collation

• Better integration of bio-informatics and cheminformatics:
  – Improved methods for prediction of potential targets and off-targets.
  – target-hopping
  – phenotypic screening

• Better integration of informatics tools into chemistry workflows

• Help chemists manage their own pitfalls.
SEA: Predicting activity via chemical similarity

Predictions derived from analysis of ChemBL database

Tremendous potential value for phenotypic screening

Help chemists Avoid Pitfalls

Computational approaches can help chemists to avoid pitfalls:

- **Over-interpretation of statistically insignificant SAR**
  - Too few datapoints, insignificant data differences.
  - Assist chemist to design experiments to enhance robustness.

- **Tendency to form SAR assumptions, and not challenge them sufficiently.**
  - “There’s no way an amine would be tolerated in that location…”
  - What is the basis of the assumption? Is it valid? How would it best be tested?

- **SAR “white-space” exploration is not usually done systematically.**
Thank you for your attention!!

Enjoy the rest of the symposium