

Light Speed MMPA and UMAP using RDKit and WebGPU

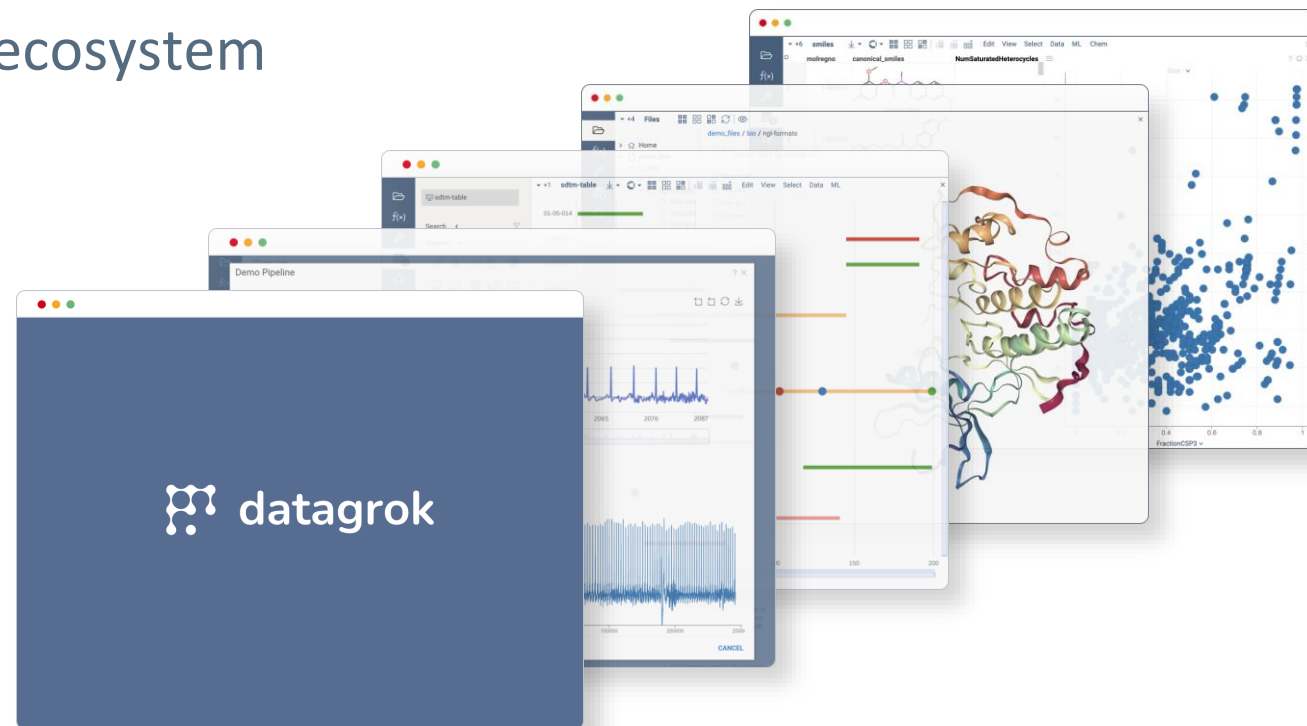
Davit Rizhinashvili, Datagrok

RDKit UGM 2024

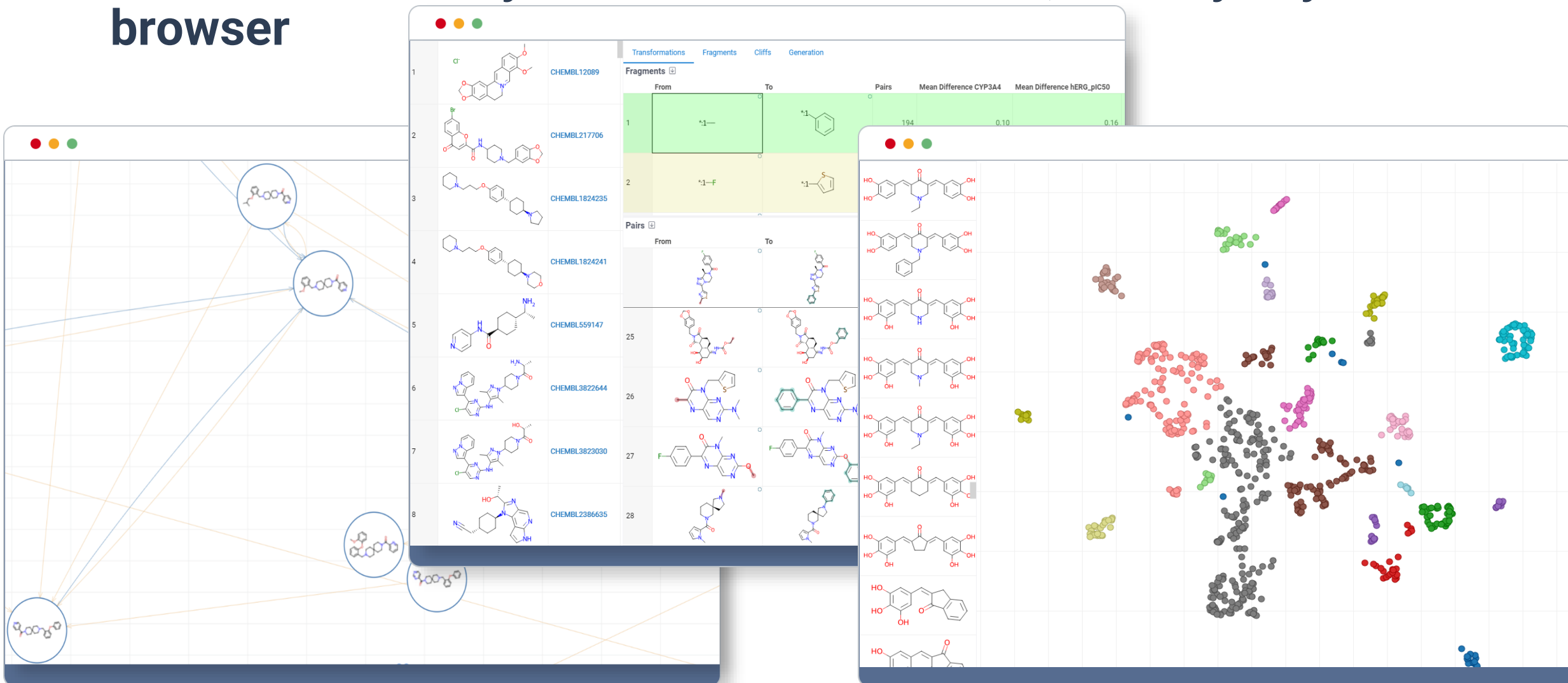
September 13, 2021

Datagrok: enterprise-ready life sciences platform

- Data access, exploratory data analysis, scientific computing, etc
- Analyzing big datasets completely in the browser
- Proprietary core, open-source plugin ecosystem
- Industry adoption
- Domain-agnostic
- Cheminformatics as a plugin
- RDKit (WebAssembly or Python)



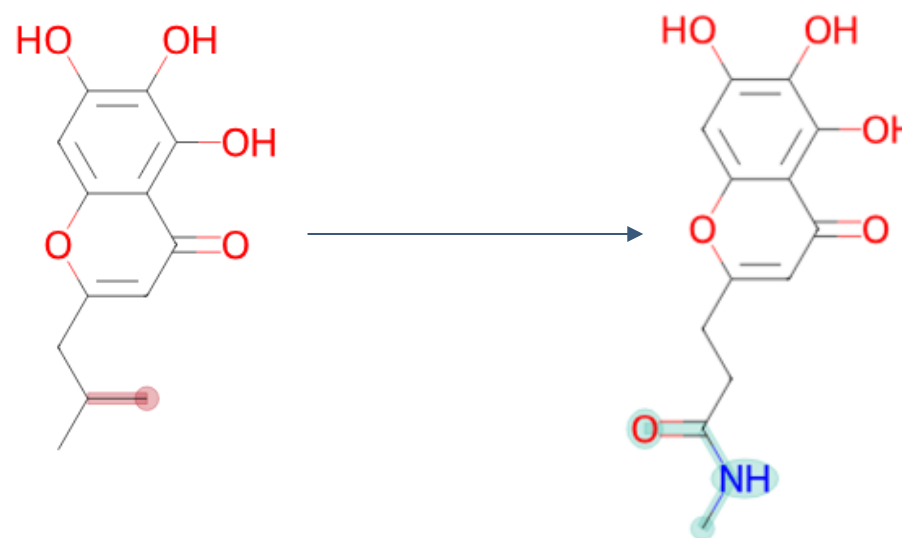
Today, we'll explore how we've achieved light speed acceleration in key cheminformatics tasks, directly in your browser



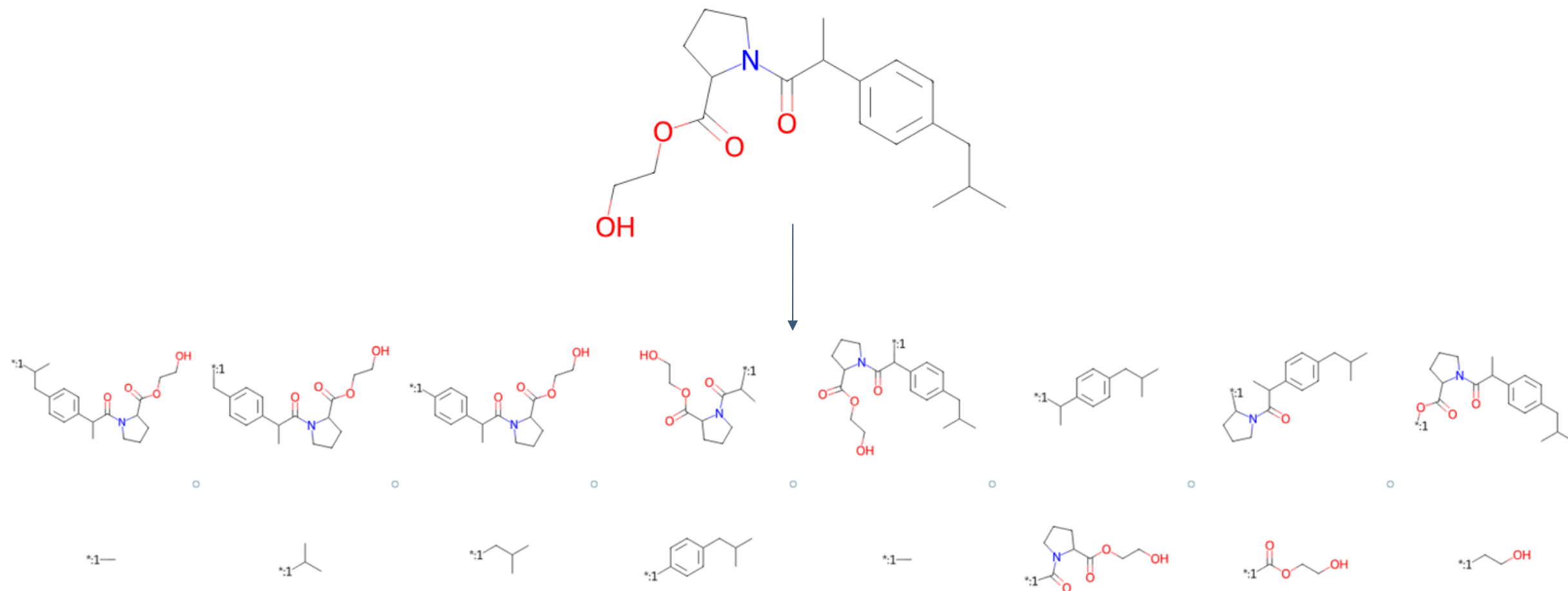
MMPA

- Fragment generation
- Identification of matched molecular pairs
- Transformation extraction
- Property change calculation
- Statistical analysis, visualization and interpretation
- New compound generation

Insights gained from MMPA can be applied to guide the design of new molecules with improved properties and predict the effects of proposed structural modifications on molecular properties

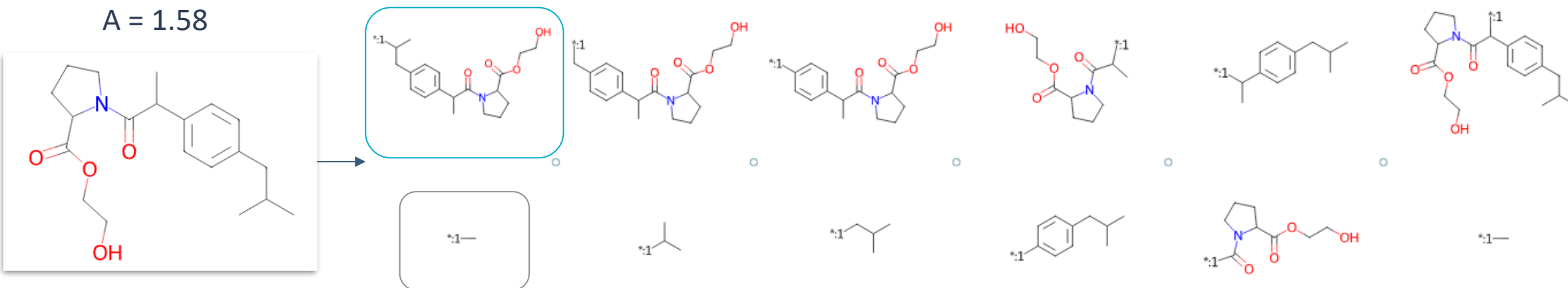


Fragmentation

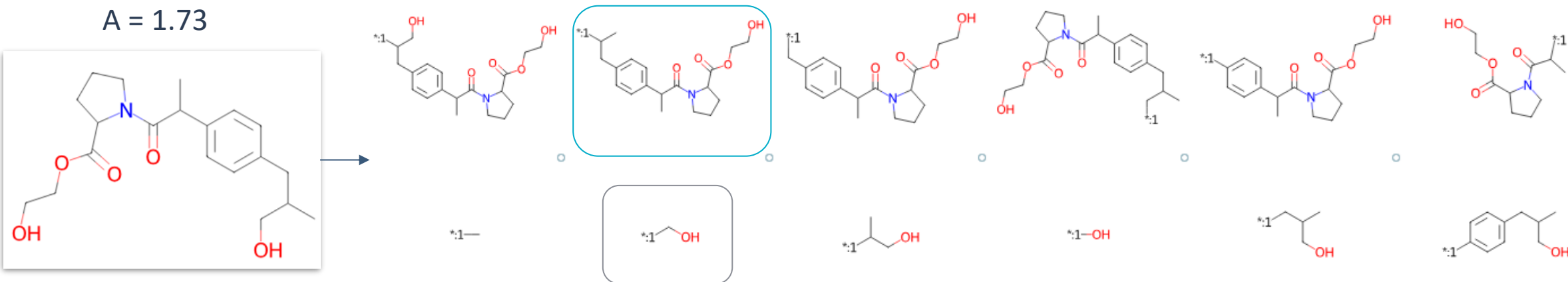


Transformations

A = 1.58



A = 1.73



Terrible scaling

For a dataset with 10_000 molecules and on average 10 fragment pairs each, transformations step would perform

$$10_000 * 10_000 * 10 * 10 = 10_000_000_000$$

complex comparisons.

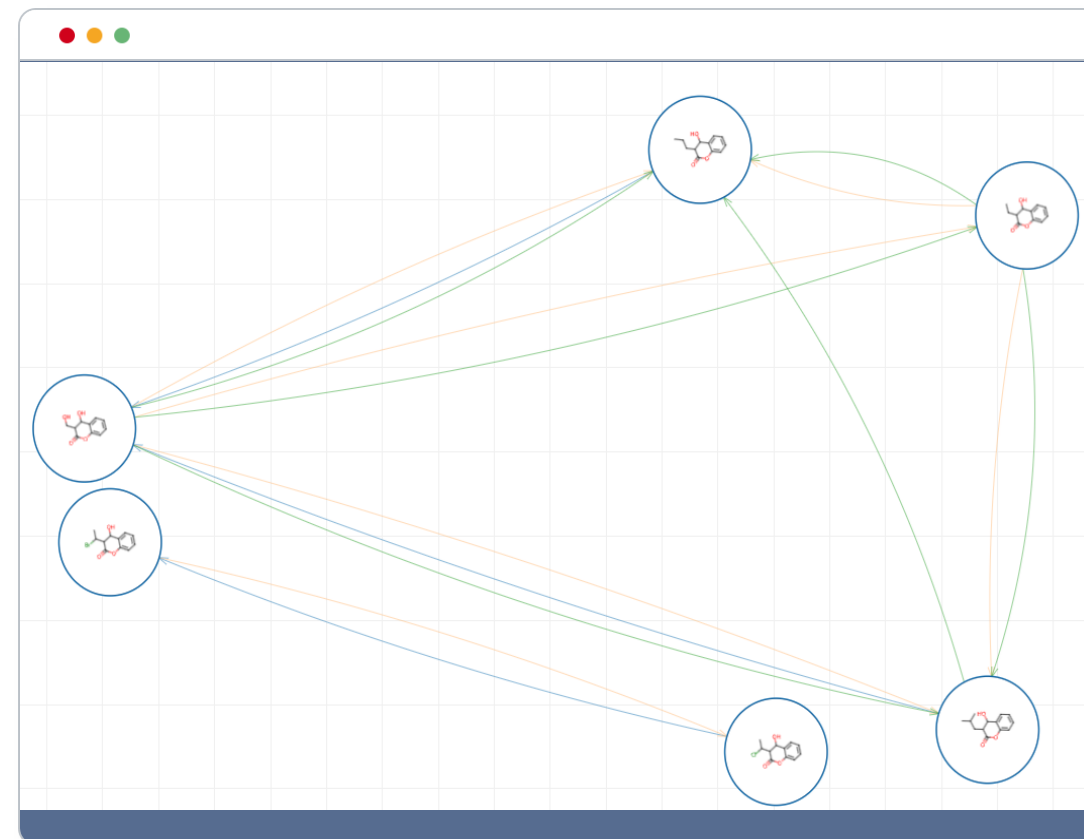
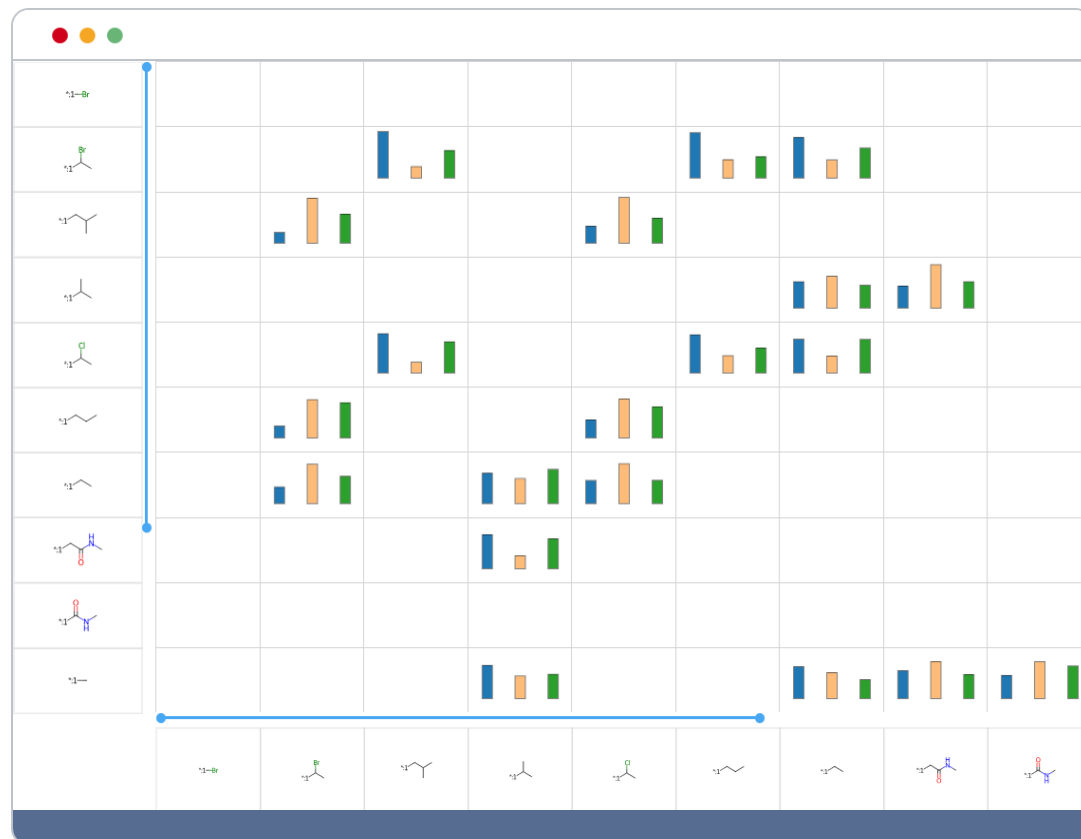
That is a lot of zeros, without even considering subsequent steps.



Analysis and visualization

- Statistics for each pair substitution
- Chemical space and activity cliffs

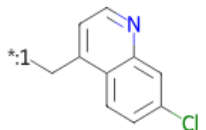


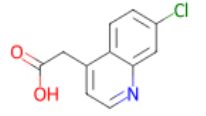
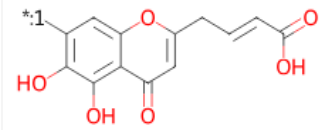


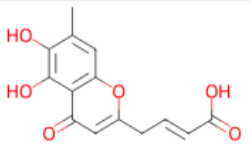
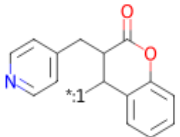


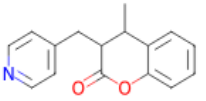
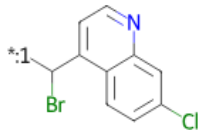


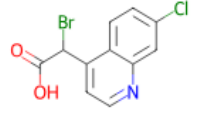
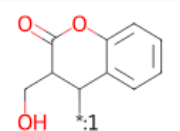


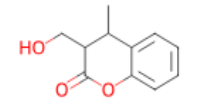
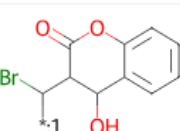


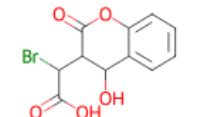
Bottleneck N2



New compound generation

Generate molecules based on transformation rules, with substitutions that yield highest activity.

Bottleneck N3

Initial value	Core	From	To	Prediction	Generation
48.67				58.98	
33.83				29.25	
36.07				31.49	
58.31				68.62	
41.26				36.69	
48.65				58.96	

Traditional methods are too slow!

Overnight jobs – BOO!

RDKit Wasm and WebGPU to the rescue!

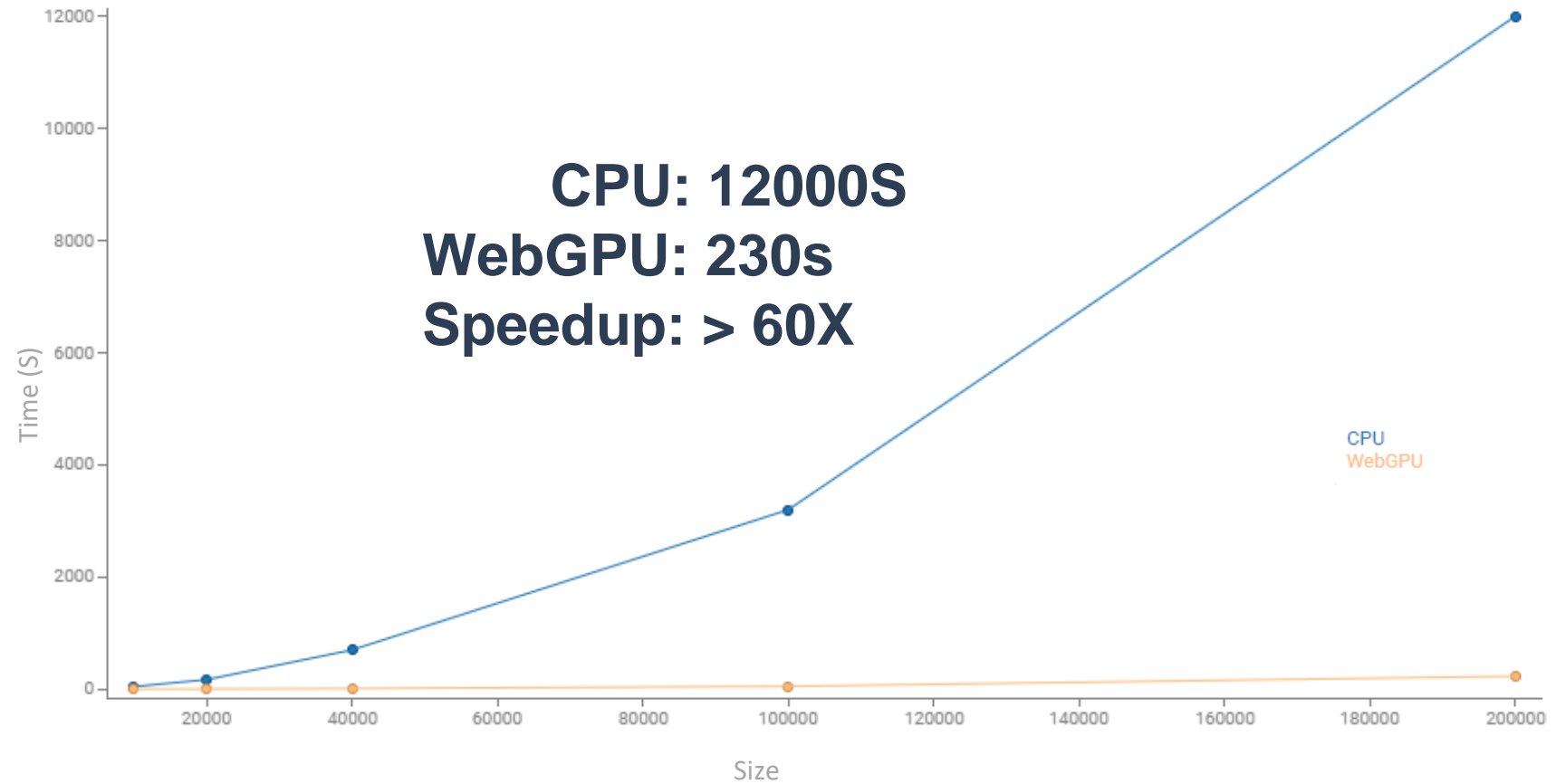
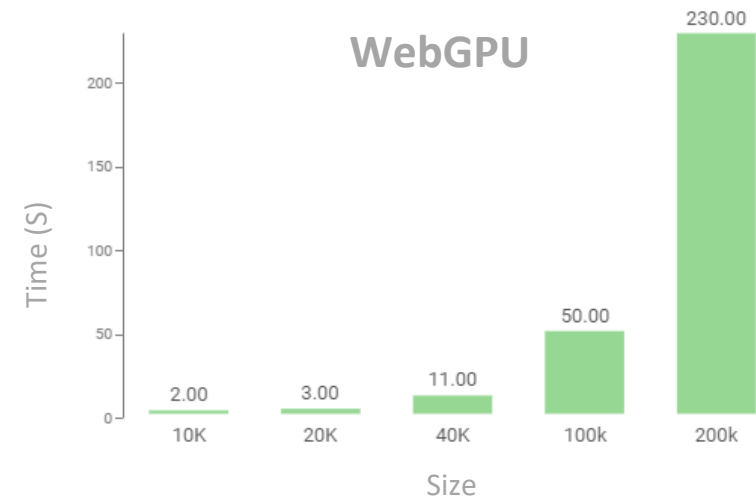
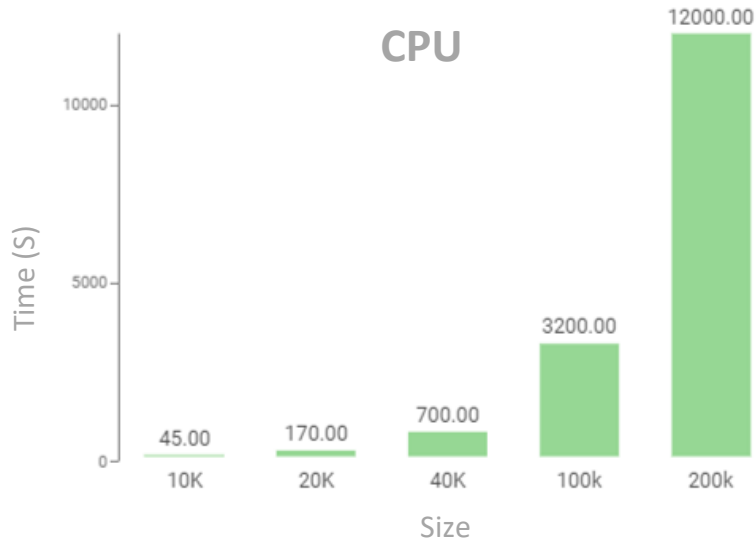


Our Approach

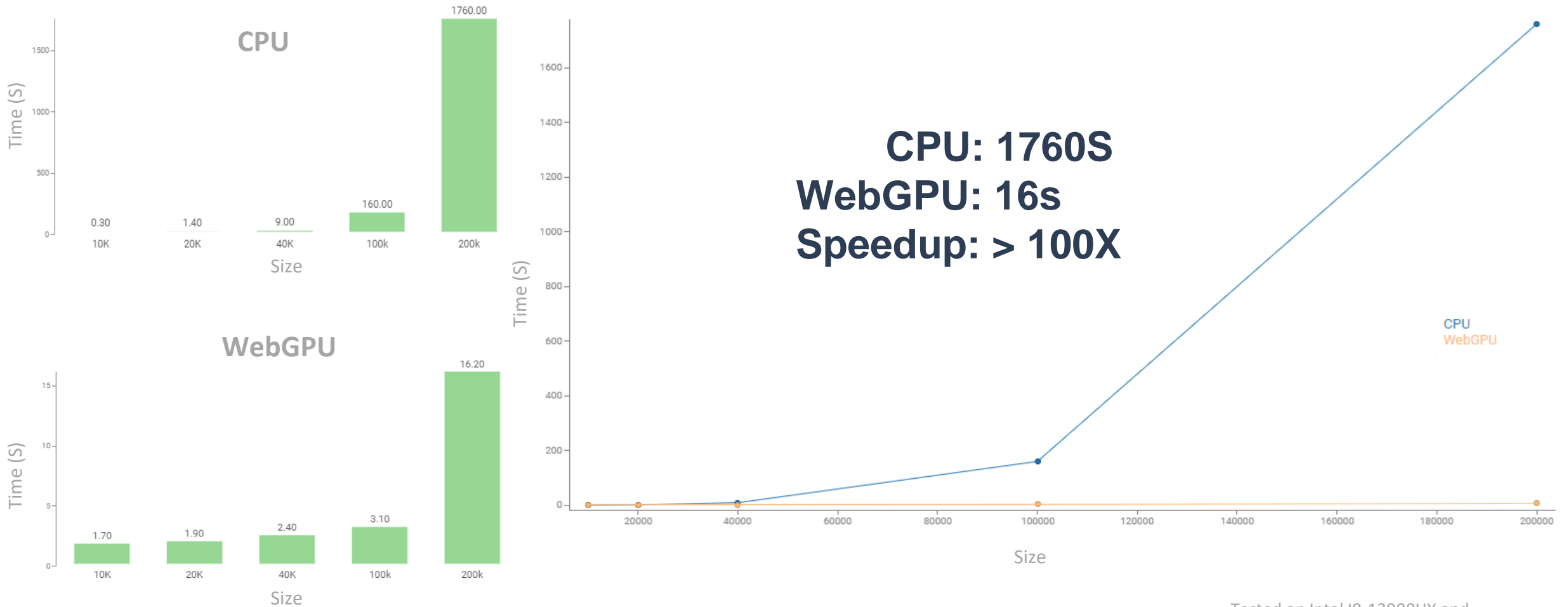
- RDKit WASM in multiple web-workers, Parallel fragmentation.
- Transformation rules calculated on WebGPU, based on encoded fragments.
- Chemical space on WebGPU, including pairwise tanimoto distances and UMAP.
- Generation of new compounds on WebGPU
- Datagrok visualization tools
- Everything in browser

Demo time!

Benchmarks: Transformation rules



Benchmarks: Generation



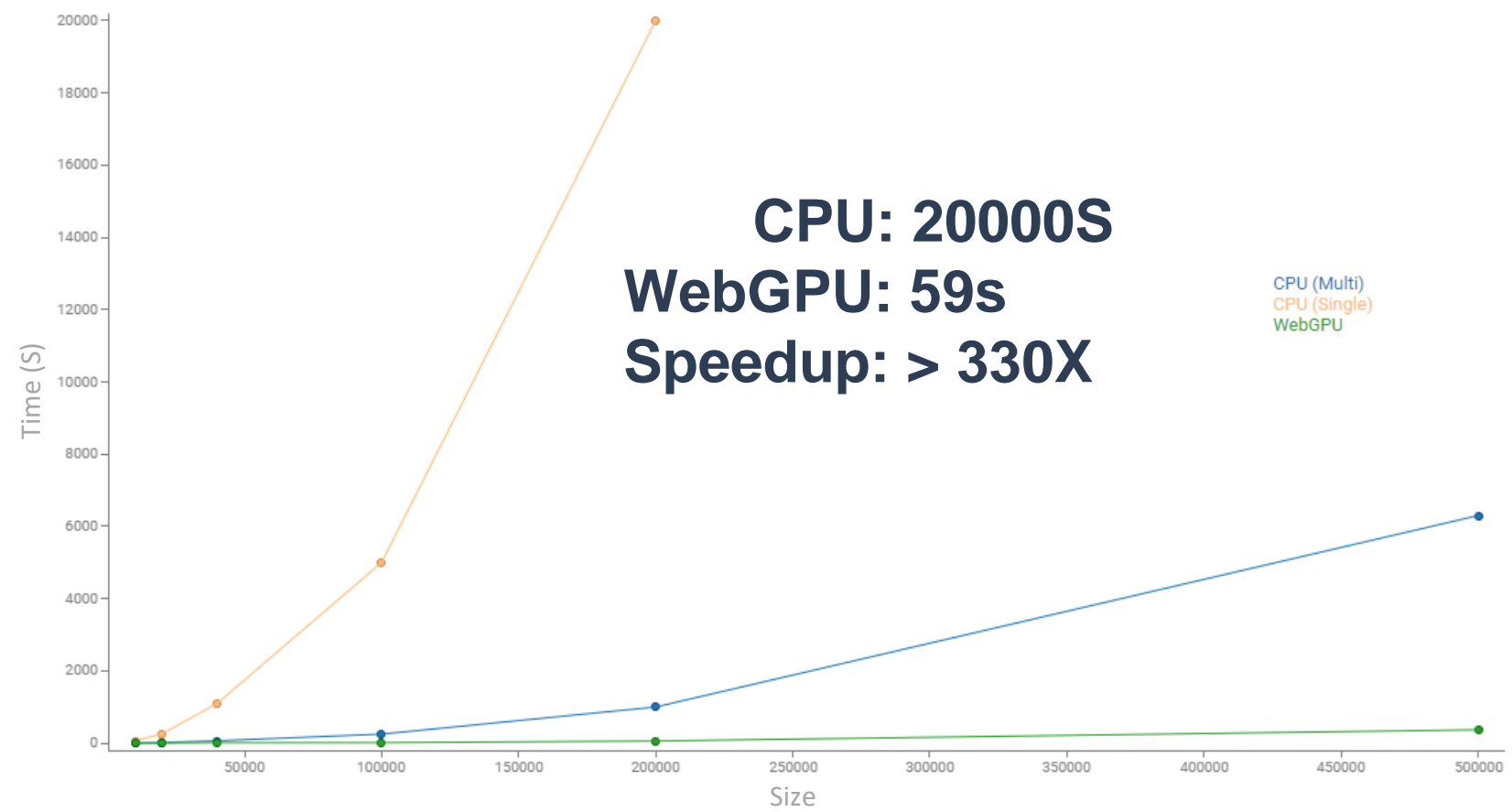
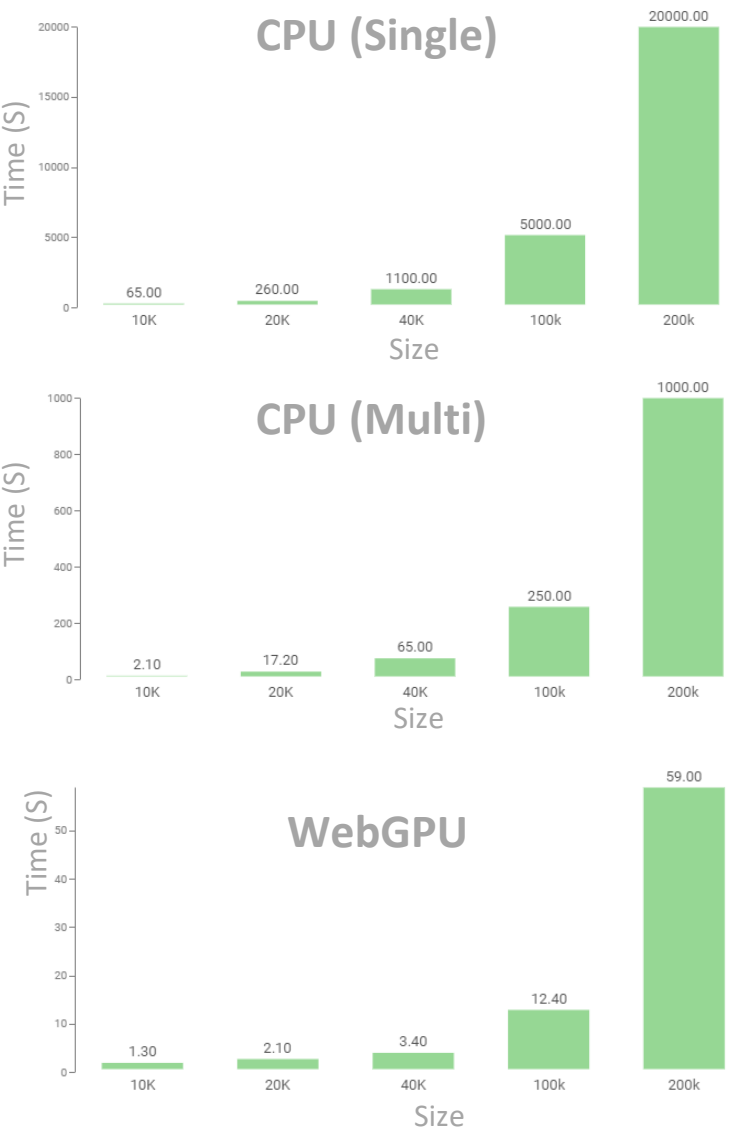
Light-speed dimensionality reduction

- Pairwise calculations performed on WebGPU, supporting complex distance functions, such as Tanimoto, Cosine, Sokal, Asymmetric for molecular fingerprints, Hamming, Levenstein, Needleman-Wunsch, Monomer chemical distance for macromolecules, vector and string operations.
- UMAP performed on WebGPU
- Up to 100X speedup, compared to multi-threaded variants.

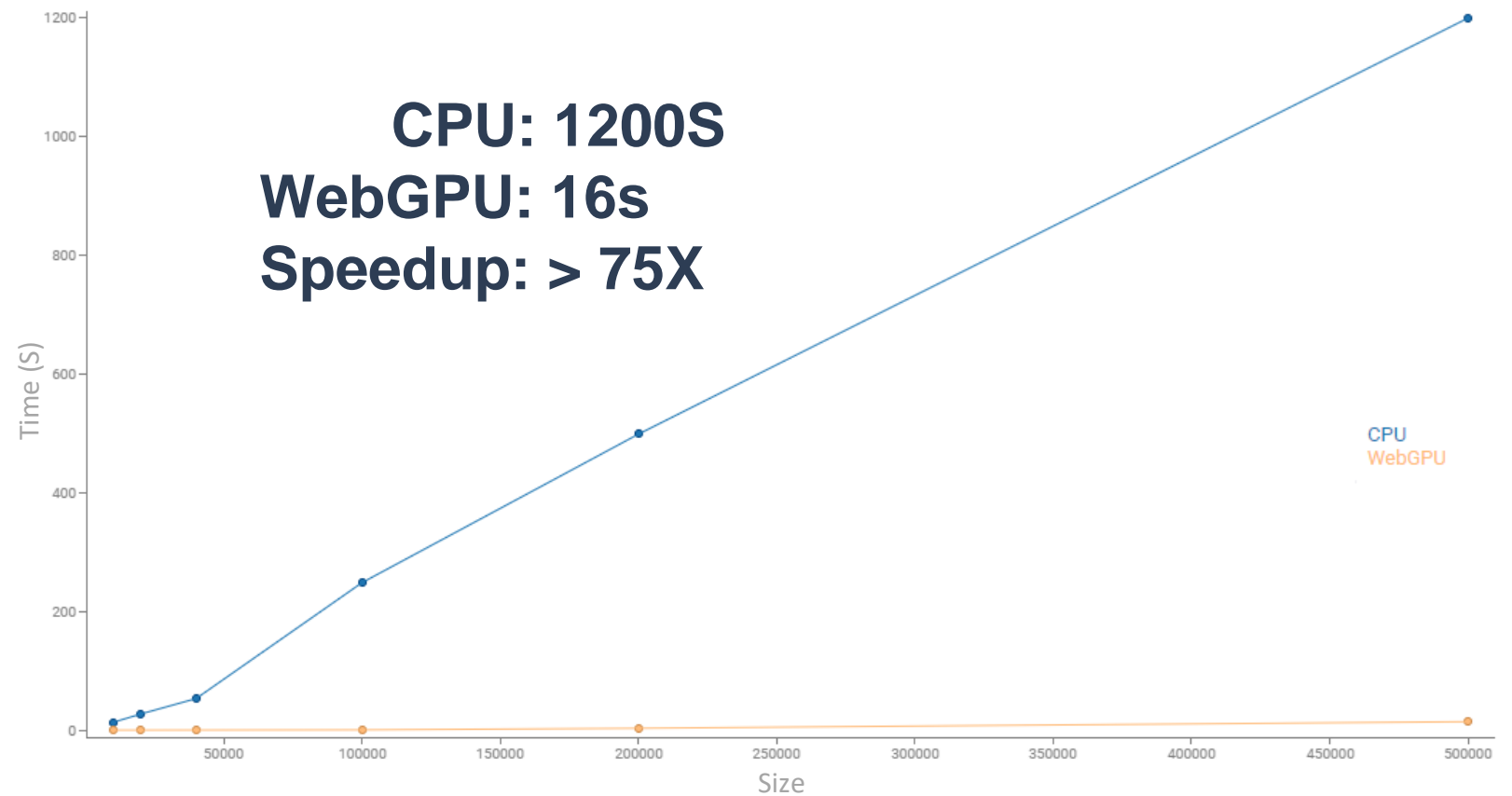
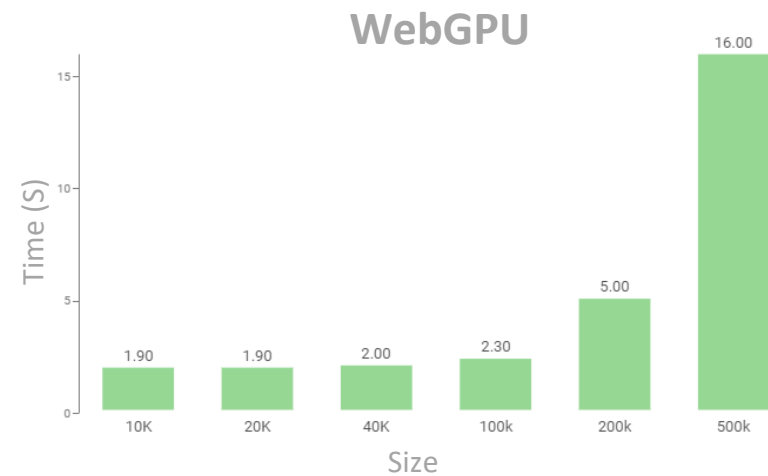
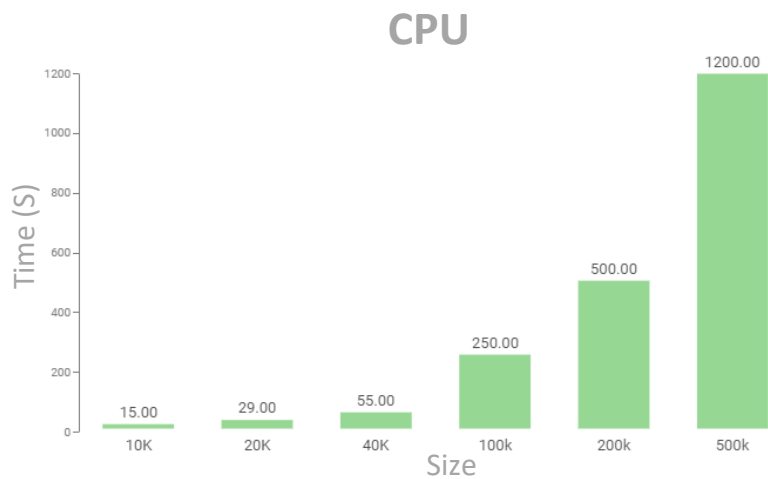
Demo Time

Benchmarks: KNN calculation

2048 bit Morgan fingerprints, Tanimoto distance



Benchmarks: UMAP



Thank you!

Acknowledgements

- Greg Landrum
- Paolo Tosco
- RDKit community
- Novartis Institutes for BioMedical Research
- Datagrok team
- All our users 😊

Run the platform right now in your browser: <https://public.datagrok.ai>

ONE MORE THING...

Datagrok is now offering free licenses to academic institutions for teaching and non-commercial research purposes!

<https://datagrok.ai>