Rethinking traditional challenges of cheminformatics and QSAR modeling in the age of ultra-large chemical libraries

Alexander Tropsha[1]

[1]*Laboratory for Molecular Modeling, UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC 27599, USA.*

Progressive advances in high throughput chemical synthesis and biological screening technologies have led to the "Bing Bang" expansion of both chemical bioactivity databases and the purchasable chemical libraries used for virtual screening. The former databases currently comprise billions of datapoints that can be used for cheminformatics model training whereas the latter libraries beginning to reach trillions of molecules. This unbelievable expansion of data notwithstanding, one may argue that major challenges of cheminformatics such as chemical similarity searching, QSAR modeling, ligand- and structure-based virtual screening, data visualization, and rational design of new chemical entities with desired properties have remained unchanged. What has changed, however, are the complexity of computational methods and tools for data representation and modeling, demands for computer power to support new methods for data processing, and incorporation of experimental validation of computational predictions as key metrics of modeling success. I will review, with examples, historical but always contemporary, Big Chemical Data informed methods for primary data curation, similarity searching, QSAR model development, validation, and out of domain extrapolation, AI-accelerated virtual screening, and both computational, and increasingly, real time active learning as part of DMTA cycle. Examples include using LLMs to curate novel federally approved medications, indications, and contraindications (MEDIC) database (*1*), QSAR modeling of multi-billion DNA encoded libraries (DELs) (*2*), tools for chemical similarity searching (*3*) and virtual screening (*4*) of trillion-size make-on-demand libraries, hit to lead optimization approaches (*5*), and property filters (*6*) to improve drug-like hit rates. I will emphasize the importance of confidence and reliable extrapolation (*7*) to ensure accuracy of machine learning based models for predicting chemical bioactivity/property. Methods and tools discussed in this presentation contribute to the overarching movement in cheminformatics in support of democratizing computational drug discovery (*8*).

Bibliography:

1. M. DeLuca *et al.*, *Nucleic Acids Res.* **54**, D1477–D1487 (2026).
2. J. Wellnitz *et al.*, *J. Med. Chem.* **68**, 21635–21648 (2025).
3. K. E. Kirchoff *et al.*, in *Advances in Information Retrieval*, N. Goharian et al., Eds. (Springer Nature Switzerland, Cham, 2024), pp. 34–49.
4. K. I. Popov, J. Wellnitz, T. Maxfield, A. Tropsha, *Mol. Inform.* **43**, e202300207 (2024).
5. HEALER: Hit expansion to advanced leads using enumerated reactions | Poster Board #1010, (available at https://acs.digitellinc.com/p/s/healer-hit-expansion-to-advanced-leads-using-enumerated-reactions-poster-board-1010-617880).
6. J. Wellnitz *et al.*, *J. Chem. Inf. Model.* **64**, 4387–4391 (2024).
7. Y. Qu, J. Wellnitz, A. Tropsha, J. Oliva, EMOE: Expansive Matching of Experts for Robust Uncertainty Based Rejection (2024), , doi:10.48550/arXiv.2406.01825.
8. A. Tropsha, H.-J. Martin, A. Cherkasov, *Drug Discov. Today*. **30**, 104341 (2025).