J Comput Aided Mol Des (2012) 26:35–38
DOI 10.1007/s10822-011-9495-0

PERSPECTIVE

# The inevitable QSAR renaissance

**Richard D. Cramer**

Received: 7 November 2011 / Accepted: 11 November 2011 / Published online: 30 November 2011


**Abstract** QSAR approaches, including recent advances in 3D-QSAR, are advantageous during the lead optimization phase of drug discovery and complementary with bioinformatics and growing data accessibility. Hints for future QSAR practitioners are also offered.

**Keywords** QSAR · CoMFA · Lead optimization

Considering the many recent publications deploring various inadequacies of current QSAR practices[1] and outcomes [1–8], as well as QSAR's relative antiquity, the reader is forgiven for any skeptical reaction to the title. Yet several CADD groups at prestigious major and medium pharmas whom I have recently encountered began discussions by stating their intent to increase their proportion of QSAR activities (while decreasing structure-based design). What is going on?

The major cause of this sudden renewal of interest in QSAR seems to be a major yet poorly met need—practical CADD guidance for lead optimization. As the longest, most expensive, and success-determining phase in most drug discovery projects [9],[2] one might suppose lead optimization to have special need for guidance from CADD. Yet current CADD methodology development activities tend much more to address the earlier hit discovery and hit-to-lead phases. One reason may be the much higher IP-based barrier between methodology development and usage experience during lead optimization. The data sets needed to validate any new methodology's value in practical LO situations are far less available.

But probably a greater reason is that lead optimization also poses a couple of demanding challenges to CADD methodologies. The first is the nature of the candidate structures. Earlier in drug discovery, the candidate structures are rather dissimilar from one another, typically exhibiting at least three orders of magnitude of differences among experimental potencies. However, during lead optimization, the candidates are much more similar, usually varying by only one or two R-groups attached to a shared core. Accordingly the variation in measured potencies is much less, seldom much more than an order of magnitude. The second challenge is the much greater number and variety of the biological observables during lead optimization, all of which need to be considered if a ranking of candidates is to be meaningful. At the same time the pace of compound synthesis is much greater. A synthetic chemist whose career is currently dedicated to a lead optimization project does not idly await lengthy computations.

Furthermore, opportunities for usefully applying QSAR approaches across all phases of drug discovery are also

R. D. Cramer (✉)
Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144, USA
e-mail: cramer@tripos.com

[1] What is meant here by "QSAR"? In agreement with Martin [34, p. 1] the distinctive characteristic of QSAR is its emphasis on biological observations as the basis for CADD activities, in contrast with the emphasis that (receptor) structure-based CADD places on physics-based models. (Of course neither approach altogether ignores the other's focus!) Is ligand similarity then a branch of QSAR (or vice versa)? Both approaches do emphasize biological observations for making potency predictions, although they differently seek either sufficiency or improvement in those potencies. (Pharmacophore approaches, being ligand-based but structurally focused, seem a third class of CADD methodologies).

[2] Lead optimization costs, per new drug introduction, are the highest of all, exceeding those of Phase II and III development because, being earlier, they generate more dead-ends and tie up capital for longer. More specifically, lead optimization accounts for 17% of total R&D cost and around 50% of discovery cost, and may be the 3rd largest opportunity area for overall R&D cost reduction.

mushrooming, thanks to the plummeting costs of acquiring and recording data points in multiplexed experiments, our rapidly expanding appreciation of biological complexity, and the inherent methodological symbioses between QSAR and bioinformatics. To cite a current example of this symbiosis, the most advanced approaches to mission-critical "off-target predictions" [10] routinely consider similarities in both ligand and biological pathway properties. Will we start regarding QSAR simply as that branch of bioinformatics that addresses chemical structure selection, in particular among ligands?

Promising methodological advances may also be a factor in this renewed QSAR interest. While the pA50 [11] predictions of structurally "local" 3D-QSAR models seem relatively reliable [2, 12],[3] sufficient to help guide lead optimization, the tedious and unavoidably subjective aspects of aligning each structure to be predicted has severely limited 3D-QSAR's practical applicability. However, the new method of topomer CoMFA (whereby rule-generated alignments are applied individually to fragments rather than complete structures) [12] is surely among the simplest, quickest, and most objective tools in today's CADD kit. And fortunately, the almost unprecedentedly accurate predictions so far reported in lead optimization projects using "topomer CoMFA", specifically a standard deviation of 0.6 between predicted and found pA50's, over 144 made-and-tested compounds from four different organizations [13–16], if continued, should further encourage its widespread application. The exceptional speed and objectivity of the topomer CoMFA protocol is also inspiring new methodological opportunities, such as virtual screening for R-groups (see footnote 3), with hits being accompanied by potency predictions, and "QSEA" [15, 17], which simplifies exploration of a so far oft-neglected issue, how a specific QSAR varies with its training set composition. Also emergent at this writing is "template CoMFA" [18], which provides the CADD expert with control over the conformation(s) used to generate a 3D-QSAR, for example in the form of a receptor-bound conformation, while retaining the desirable attributes of topomer CoMFA.

Yet another massive development that favors usage of QSAR, with its need for experimental results to drive its hypothesis generation, is the growing capabilities of most drug discovery organizations for making experimental data, both public [19] and private, completely and readily available, and in as many comparative formats as possible.

But what about those publications deploring QSAR deficiencies? For the most part it is agreed that these disappointments are caused more by faulty practice or unrealistic expectations than by fundamental deficiencies in the QSAR approach. However it would be much harder to reach a consensus on which practices are faulty, and the recurring empirical dilemma—is my latest QSAR prediction trustworthy enough to guide a critical project decision?—remains problematic. With these caveats, here are some hints for the future QSAR practitioner, based on over 40 years of experience.

- A QSAR is simply a coherent structurally-based summary of a particular set of biological observations, hopefully revealing a pattern which helps to successfully guide a discovery team to a therapeutic goal. Yet the still barely understood complexities of biological processes, at molecular, cellular, and organism levels, surely set unknowable boundaries—aka "activity cliffs" [20]—on the extrapolability of any QSAR. (And the highly multidimensional character of most QSARs dictates that almost any worthwhile prediction is an extrapolation [21].)
- Nevertheless, it is an empirical fact, as indicated above, that QSAR predictions are often accurate enough to benefit discovery. I believe this tendency to be evidence for undiscovered regularities in biological phenomena that physical and systems biology modeling will eventually reveal. Yet meanwhile—if no drugs are found, the entire discovery process is jeopardized. If QSAR may help and is conveniently available, shouldn't it be tried?
- Each discovery team member is already considering the same observations to seek the same goal, often generating intuitive SARs similar to the current QSAR. However, the now overwhelming quantity of such observations suggests that a QSAR exercise may also be increasingly useful in calling attention to outliers and/or activity cliffs, to be scrutinized with the hope of exploitation. Having available such a single and relatively formal QSAR expression of the team's intuitive and varied SAR models makes the recognition of an outlier more likely.
- Any particular QSAR is probably only one of many statistically acceptable and perhaps equally plausible alternative QSARs, considering for example the thousands of possibly explanatory structural descriptors that are available. Rapid detection and inclusion of an "activity cliff" observation, though perhaps immediately disappointing, may highlight a more productive QSAR hypothesis.
- Statistical parameters have marginal relevance when assessing the soundness of a QSAR prediction. For example, choosing as "the model" simply the one having the greatest $q^2$ (or $r^2$) value is little better than a coin flip, considering the other uncertainties discussed in previous hints.

[3] "pA50" is used as an abbreviation for the $\log_{10}$ of any biological potency measurement.

- Also underappreciated is the strong dependence of $q^2/r^2$ on the spread in the underlying biological potencies. More informative is the entirely model –dependent standard error in the potency predictions/fits which accompanies a $q^2/r^2$. This consideration further implies that for QSAR derivation a wide spread among the biological observations, though desirable, is far from mandatory.
- Furthermore, excessive dependence on $q^2$ can seriously impede the discovery of useful new hypotheses, particularly when first encountering an "activity cliff". The $q^2$ "leave-out-and-predict" philosophy necessarily suppresses any new hypothesis that is supported by only one newly tested structure. And it would seem, the higher a cliff, the more likely its $q^2$ rejection.
- "Leave-one-out" cross-validation is also a misnomer in most QSAR derivations, because most structural series contain multiple instances of the more successful R-groups, whose influence therefore cannot be erased by omitting and predicting individual structures (see footnote 3).
- Discovery projects seek better structures. Identifying a structure which is then found to be superior, even if its potency prediction is numerically relatively inaccurate, is far more valuable than many accurate pA50 predictions of potencies already achieved (see footnote 3). Therefore judicious extrapolation of a QSAR seems desirable rather than "dangerous".
- In general classical statistics is far too optimistic when validating a QSAR, because its underlying assumptions about data distributions are contradicted by the extraordinarily heterogeneous nature of chemical structures and mechanisms of biological response. Restricting the structural scope of a QSAR should help, but the distribution of "local" structural variations, within a series undergoing lead optimization, is also unlikely to be uniform.
- One tactic to consider for prediction validation is to seek multiple QSAR models encompassing different "radii" of structural variation, with the goal of detecting as many activity cliffs as possible. A prediction that is reproduced by such a varying scope of QSARs is more likely to be robust.
- More sophisticated means of "data mining"—neural nets, pattern recognition, support vector machines, even mere non-linear regression—have been repeatedly tried [22, 23] and abandoned, as too costly and/or unreliable. The major exceptions, PLS [24], cross-validation/bootstrapping [25], and recursive partitioning [26], are relatively minor extensions of QSAR's original multiple linear regression [27–29]. It seems that biological data are too fuzzy and alternative explanatory hypotheses too numerous for QSAR to benefit from increased model complexity.
- When selecting candidate explanatory structural descriptors for use in a QSAR:
  - Physics-based descriptors [30], for example 3D-QSAR's fields, are likely to produce the most robust and useful models, for example capable of actively generating and selecting among quite varied structural ideas. Other classes of descriptor are more or less limited to passive discrimination among less structurally varied ideas, generated by some external process.
  - Substructural descriptors, for example "2D fingerprints", would seem the least reliable, because biological receptors are affected only by the fields and mechanical behaviors presented by a candidate ligand, not by the underlying atomic connectivities [31]. Yet, substructural descriptor similarity has been a better predictor of biological similarity than many 3D similarity metrics [32].
  - Whenever the underlying biological observables may depend on transport as well as receptor fit, such as passive membrane penetration, "1D" descriptors, such as log P, polar surface area, and pKa, should be considered. Perhaps obvious, yet for example 3D-QSAR models almost universally omit them [33].

Unfortunately QSAR's relatively long history also suggests that many of its future practitioners will be variously hampered in learning from past experiences. The motivations to "turn off the brain and turn on the computer" will not disappear and indeed are probably increasing. Thus it is also my somewhat biased suggestion that in such hampered situations topomer CoMFA may nevertheless be of some benefit to a discovery process.



## References

1. Golbraikh A, Tropsha A (2002) J Mol Graph Model 20:269–276
2. Doweyko A (2004) J Comput Aided Mol Des 18:587–596
3. Maggiora GM (2006) J Chem Inf Model 46:1535
4. Johnson SR (2006) J Chem Inf Model 48:25–26
5. Doweyko A (2008) IDrugs 11:894–899
6. Dearden JC, Cronin MTD, Kaiser KLE (2009) SAR QSAR Environ Res 20:241–266
7. Scior T, Medina-Franco JL, Do Q-T, Martinez-Mayorga K, Yunes-Rojas JA, Bernard P (2009) Curr Med Chem 16: 4297–4313
8. Czerminski R, Manchester J (2008) J Chem Inf Model 48: 1167–1173
9. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindberg SR, Schacht AL (2010) Nat Rev Drug Discov 9: 203–214

10. Abernethy D, Bai J, Burkhart K, Xie H-G, Zhichkin P (2011) Clin Pharmacol Ther 90:203–214
11. Cramer RD, Cruz P, Stahl G, Curtiss WC, Campbell B, Masek BB, Soltanshahi F (2008) J Chem Inf Model 48:2180–2195
12. Cramer RD (2003) J Med Chem 46:374–389
13. Martin YC (2011) J Comput Aided Mol Des 25:195–196
14. Cramer RD (2011) J Comput Aided Mol Des 25:197–201
15. Wendt B, Mülbaier M, Wawro S, Schultes C, Alonso J, Janssen B, Lewis J (2011) J Med Chem 54:3982–3986
16. Tresadern G, Bemporad D (2010) Fut Med Chem 2:1547–1561
17. Wendt B, Cramer RD (2008) J Comput Aided Mol Des 22:541–551
18. Cramer RD (2011) Abs ACS Fall Mtg COMP 125
19. Wendt B, Uhrig U, Boes F (2011) J Chem Inf Model 51:843–851
20. Maggiora GR (2006) J Chem Inf Model 46:1535
21. Clark RD, Cramer RD (1997) Chemtech 27:24–30
22. Cramer RD (1976) Ann Rep Med Chem 11:301–310
23. Kowalski BR, Bender CF (1973) J Am Chem Soc 95:586
24. Wold S, Ruhe A, Wold H, Dunn WJ (1984) SIAM J Soc Stat Comput 5:735
25. Cramer RD, Bunce JD, Patterson DE (1988) Quant Struct Act Relatsh 7:18–25
26. Hawkins DM, Young SS, Rusinko A (1997) Quant Struct Act Relatsh 16:296–302
27. Hansch C, Maloney PP, Fujita T, Muir RM (1962) Nature 194:178–180
28. Free SM, Wilson J (1964) J Med Chem 7:395–399
29. Hansch C, Fujita T (1964) J Am Chem Soc 86:1616–1626
30. Unger SH, Hansch C (1973) J Med Chem 16:745
31. Cramer RD, Redl G, Berkoff CE (1974) J Med Chem 17:533–535
32. Brown RD, Martin YC (1999) J Chem Inf Comput Sci 36:573–584
33. Cramer RD, Wendt B (2007) J Comput Aided Mol Des 21:23–32
34. Martin YC (2010) Quantitative drug design: a critical introduction. CRC Press, Boca Raton, p 1