Initial Evaluation of QSAR Modeling for Screening the CCL Universe to the PCCL

Report for the NDWAC CCL Work Group
Plenary Meeting
September 17-18, 2003

Purpose and Key Conclusions

- Purpose was to evaluate , as requested by NDWAC CCL WG, the use of QSAR models to provide occurrence and health effects information for diverse chemical sets for screening from the Universe to the PCCL.
- QSAR modeling appears feasible for predicting water solubility and biodegradability for a wide range of contaminants.
- QSAR modeling for chronic toxicity also appears feasible, but range of contaminants for toxicity modeling is not as broad as for physical-chemical modeling.

Overview

- Evaluated utility of QSAR programs to predict chronic toxicity, water solubility and biodegradability of a sample dataset of CCL Universe chemicals
  - Applied TOPKAT program to predict oral rat chronic toxicity, i.e., Lowest Observable Adverse Effect Level (LOAEL)
  - Applied EPI-Suite^{TM} program models to predict water solubility (WSKOWWIN) and biodegradability (BIOWIN)

QSAR Model Selection: TOPKAT

- TOPKAT: "The Open Practical Knowledge Acquisition Toolkit"
  - Commercial package licensed by Accelrys (accelrys.com/products/topkat/). Algorithms and training sets proprietary
  - Uses 2-D descriptors of chemical structural information (SMILES) to predict range of human health properties
  - Rat oral LOAEL model developed from IRIS and NTP databases of chronic toxicity values
  - Currently in use by ORD NCEA and other regulatory institutions worldwide

QSAR Model Selection: EPISUITE

- EPI-Suite^{TM}: “Estimation Program Interface-Suite” of 12 models developed by EPA OPPT and Syracuse Research Corporation
- Predicts physical/chemical properties and environmental fate measures
- Publicly available through EPA, user-friendly with SMILES input, and validated
- Used by OPPT and other regulatory institutions worldwide for chemical assessment and prioritization

QSAR Contaminant Test Set

- Two main categories of test set contaminants:
  - Those with existing empirical data for evaluating how well the models performed
  - Those without existing empirical data for evaluating applicability of models to contaminants lacking data
- Contaminant test set built from 3 Groups
  - Group 1: Draft 1998 CCL1 List, plus ~25 compounds (endocrine-disrupting compounds and pesticides) deferred from that list.
    - Group 1 had some with and some without empirical LOAEL data.
  - Group 2: Non-CCL1 chemicals with empirical LOAEL data
  - Group 3: Non-CCL1 chemicals lacking empirical LOAEL data

Compiling and Sorting Chemicals for QSAR Evaluation

❑ Compiled initial datasets	1,866 chemicals
❑ Limiting criteria:	
▪ Duplicates	- 562
▪ Regulated chemicals	- 6
▪ EPI SuiteTraining set chemicals	- 369(*)
▪ Not conducive for QSAR	- 234
❑ Final QSAR Contaminant Set	695 chemicals
▪ Group 1 (CCL1 w LOAEL):	60
▪ Group 2 (Non-CCL w/ LOAEL):	167
▪ Group 1 (CCL1 w/o LOAEL):	81
▪ Group 3 (Non-CCL w/o LOAEL):	387

(*) – An additional 21 TOPKAT training set chemicals subsequently identified.

Compiling and Sorting Chemicals for QSAR Evaluation

- Final QSAR Contaminant Set 695 chemicals
  - Availability of Empirical LOAEL Data
    - With LOAEL from CCL1 (Group 1): 60
    - With LOAEL from Non-CCL (Group 2): 167
    - Without LOAEL from CCL1 (Group 1): 81
    - Without LOAEL from Non-CCL (Group 3): 387
  - Availability of Empirical Solubility Data
    - With water solubility data: 296

      Group 1: 33% Group 2: 54% Group 3: 41%
    - Without water solubility data: 399
  - Biodegradation Data Available for All

Empirical Data Sources: Chronic Toxicity Data

- Registry of Toxic Effects of Chemical Substances (RTECS)
  - Rat or mouse oral LOAELs from studies of 28 days or longer (TD_{lo}’s in RTECS)
    - Tried to limit to >90 day rat, but too few studies available
  - 892 LOAELs extracted for 227 chemicals (156 with multiple values)
  - Cumulative doses reported by RTECS converted to daily dose (mg/kg-day)

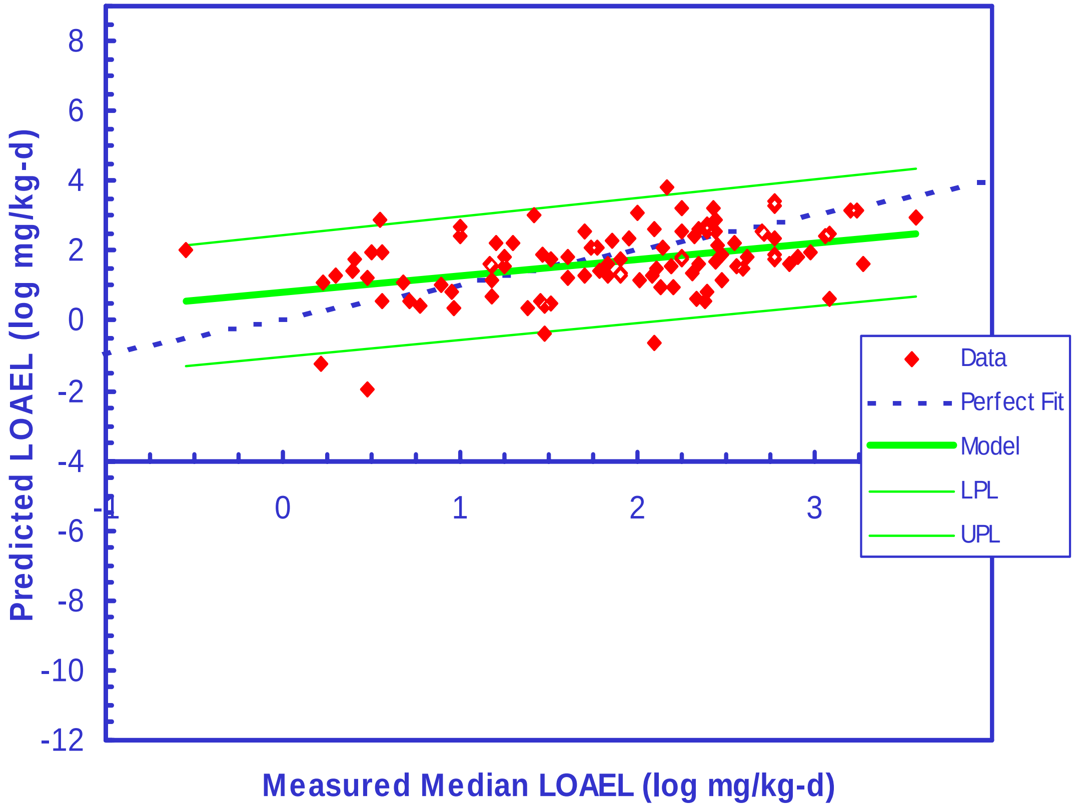
Empirical Data Sources:

Solubility and Biodegradability

Data Source	Number of Chemicals with Empirical Data
SRC -CHEMFATE and BIODEG databases, Syracuse Research Corporation (http://esc.syrres.com/efdb.htm)	132
HSDB - Hazardous Substances Data Bank (http://toxnet.nlm.nih.gov/)	109
MacKay - MacKay, Shiu, and Ma 1999 (Physical-Chemical Properties and Environmental Fate Handbook)	46
NTP - National Toxicology Program (http://ntp-server.niehs.nih.gov/)	32
IPCS - International Program of Chemical Safety (http://www.inchem.org/)	1

TOPKAT Predictions Versus Measured LOAELs

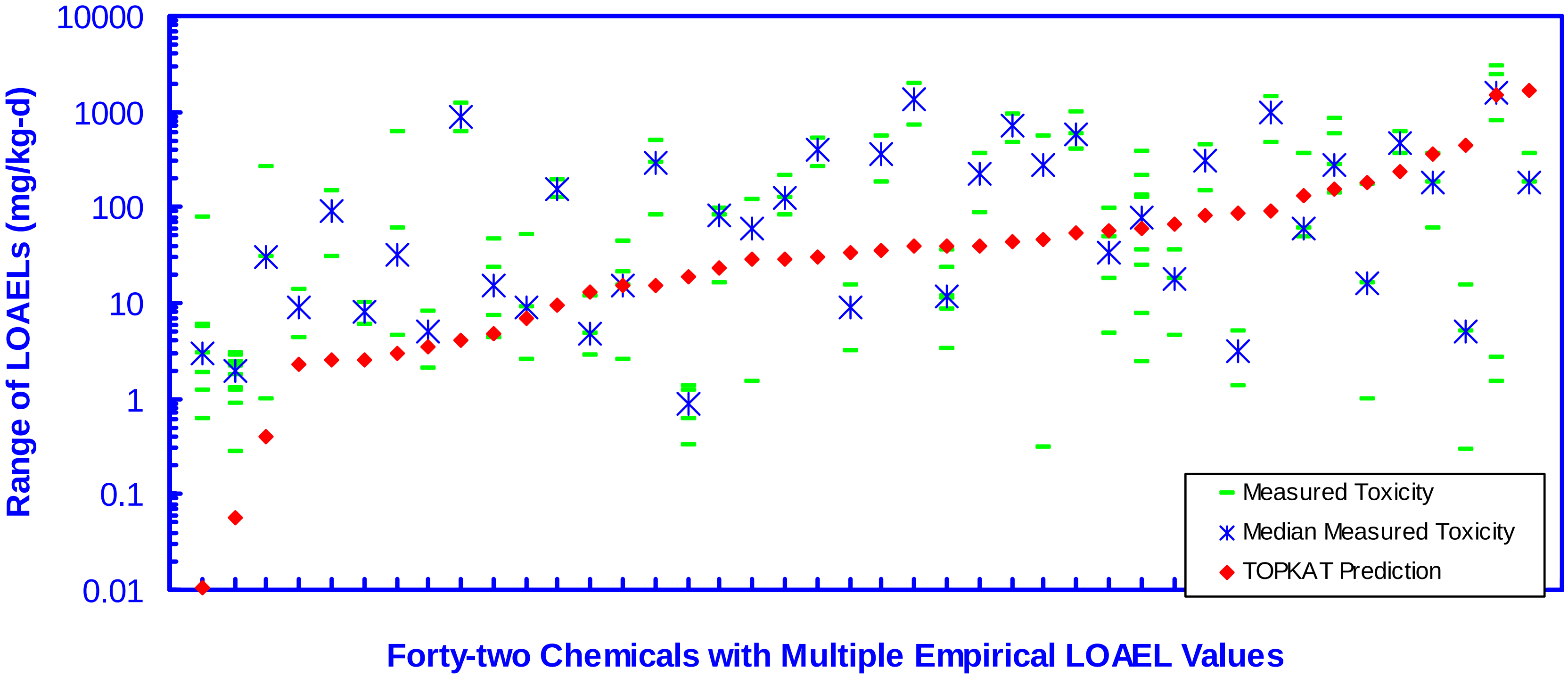
Data from 28-day minimum duration, rat or mouse bioassays; predictions with TOPKAT error codes excluded from analysis.



⇨ Significant correlation for chemicals within TOPKAT's predictive domain

Comparison of TOPKAT Predictions for Chemicals with Multiple Measured LOAEL Values

Measured values from rat bioassays ≥90-days; predictions with error codes excluded. (Chemicals ordered from most to least toxic by TOPKAT prediction.)



Range of empirical LOAELs broad and impedes model comparison to measured values.

Cross Validations of TOPKAT Results

Validation	**Percent Predictions within Factors of Empirical Estimates**				
	Factors:	2	5	10	100
Accelrys Reported (Goodness-of-Fit; Training Set)	All Models (averaged)	76	99		
Mumtaz et al. 1995 Training Set	All Models	55	94	100	
NCEA Preliminary Data Separate Test Set	All Models, chronic duration, rat only, no error-coded data	**33**	**60**	**72**	**98**
Cadmus Test Sets	All Models, ≥28 duration, rat and mouse data, no error-coded data	**20**	**53**	**68**	**95**
Cadmus Test Set	All Models, ≥90 day studies only	14	48	62	92

⇨ Prediction factors indicate non-training set chemicals do not perform as well as training set chemicals.

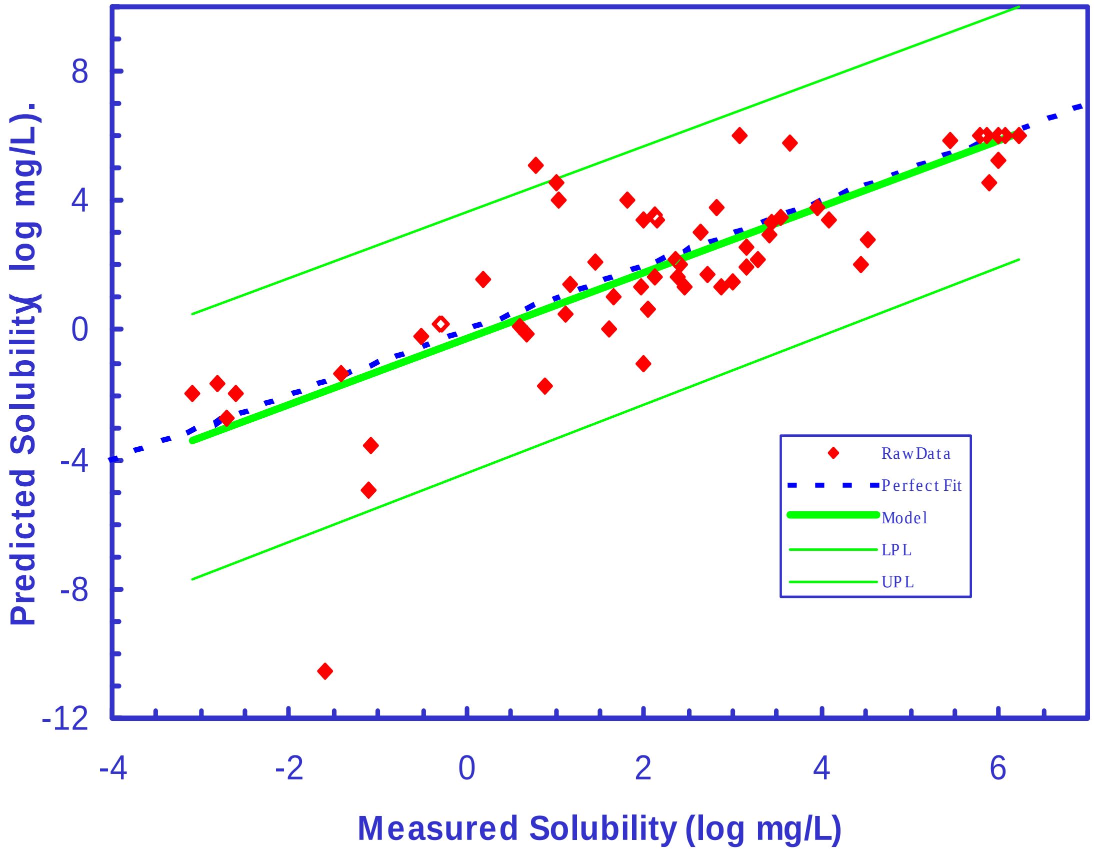
However, results similar to NCEA's preliminary validation observed in this analysis, despite study design differences. QSAR modeling for CCL Universe to PCCL appears feasible.

Conclusions: Chronic Toxicity

- TOPKAT able to predict LOAELs for 45% of chemicals tested.
- TOPKAT reliably identified those chemicals outside its domain. (55% of queries limited by model coverage.)
- Validations of TOPKAT comparable with preliminary results from NCEA, given study design differences. 72% of NCEA and ~ 65% of this effort's predictions within a factor of 10 of empirical values.
- TOPKAT results generally accepted as valid, despite limited transparency.

EPI-Suite^{TM} Predictions of Solubility

Data shown are from CCL chemicals; non-CCL chemicals yielded similar results.

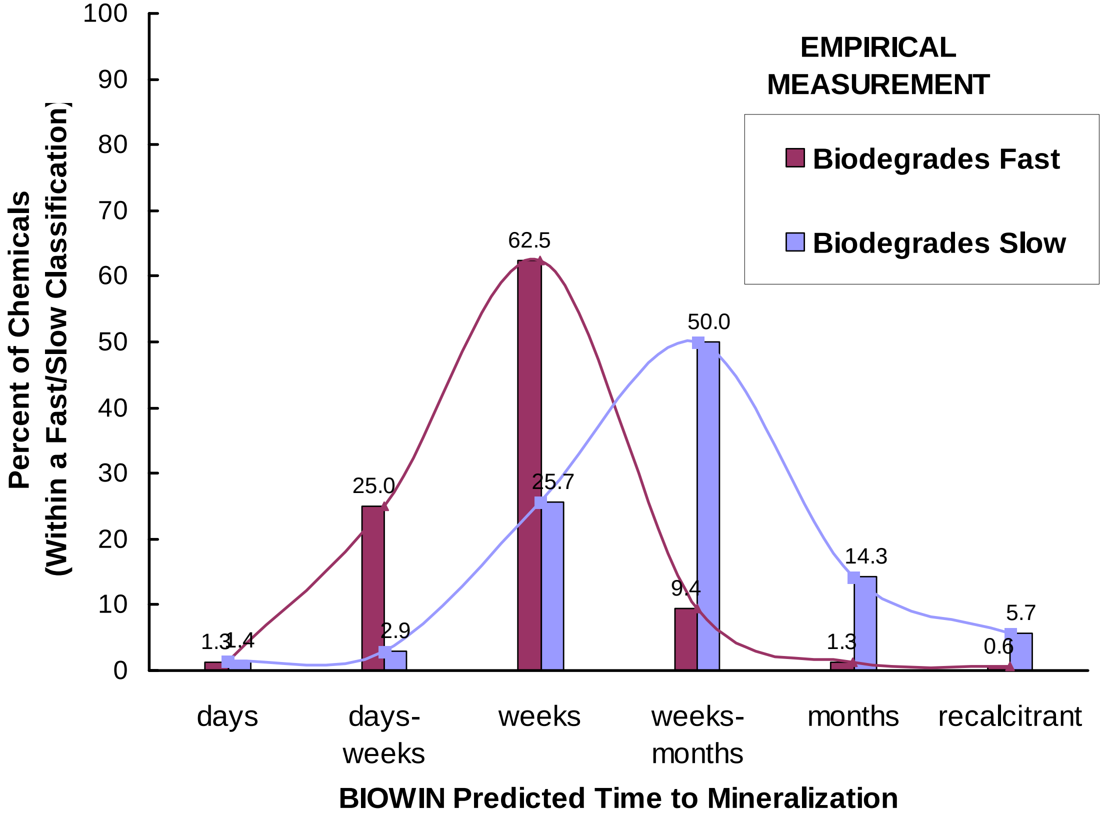


EPI Suite solubility predictions within a factor of 5 for 54% of queries. Appears feasible to use QSAR modeling for CCL Universe to PCCL.

Conclusions: Water Solubility

- ❑ WSKOWWIN predicted water solubility within a factor of five for 54% of chemicals evaluated
- ❑ Relatively high variability observed among 46 chemicals with multiple empirical values, due to methods diversity and variability
- ❑ WSKOWWIN is user-friendly, transparent, and applicable to broad range of chemicals

Results of BIOWIN Predictions of Biodegradation



BIOWIN distinguished fast and slowly degrading compounds.

QSAR modeling appears feasible for CCL Universe to PCCL screening.

Conclusions: Biodegradability

- QSAR modeling appears feasible for semi-quantitative estimation (categorization) of biodegradability
- BIOWIN predictions broadly distinguished fast- and slowly-degrading chemicals
- Empirical data from certain test procedures (e.g., ready tests) may limit comparisons of predictions of rates of mineralization (e.g., weeks, months) from BIOWIN

Overall Conclusions from Initial QSAR Modeling Evaluation

- ❑ QSAR modeling with EPI-Suite^{TM} appears feasible for predicting water solubility and biodegradability for use in CCL Universe to PCCL screening
- ❑ QSAR modeling using TOPKAT for health effects appears possible, but may require greater selectivity in chemicals and health effects modeled
- ❑ Comparison of QSAR model results to empirical data limited by missing and highly variable measurements reported. This may generally limit the ability to fully evaluate QSAR model predictions for chemicals outside their respective training sets.

Overall Conclusions from Initial QSAR Modeling Evaluation (Cont.)

- Initial QSAR modeling suggests QSARs more readily generate information on occurrence-related properties than human health effects endpoints for large and diverse chemical sets.
- Chemical input development may be more resource-intensive than actual QSAR modeling.
- Processing thousands of chemicals using QSAR models would be facilitated by efficient batch mode operations.